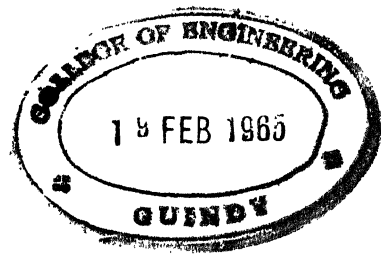# TRANSACTIONS

OF THE

AMERICAN INSTITUTE

OF

ELECTRICAL ENGINEERS

VOLUME 74

1955

PART II. APPLICATIONS AND INDUSTRY

Preface

The AIEE *Transactions* for 1955 (volume 74) is published in three parts: Part I. Communication and Electronics; Part II. Applications and Industry; and Part III. Power Apparatus and Systems. The papers in each of the three parts are classified according to subject matter as follows:

Part I. Communication and Electronics	Part II. Applications and Industry	Part III. Power Apparatus and Systems
Communication Switching Systems	Air Transportation	Carrier Current
Communication Theory	Domestic and Commercial Applica-	Insulated Conductors
Radio Communications Systems	tions	Power Generation
Special Communications Applications	Land Transportation	Protective Devices
Telegraph Systems	Marine Transportation	Relays
Television and Aural Broadcasting	Production and Application of Light	Rotating Machinery
Systems	Chemical Industry	Substations
Wire Communications Systems	Electric Heating	Switchgear
Basic Sciences	Electric Welding	System Engineering
Computing Devices	Feedback Control Systems	Transformers
Dielectrics	General Industry Applications	Transmission and Distribution
Electrical Techniques in Medicine and	Industrial Control	
Biology	Industrial Power Rectifiers	
Electronics	Industrial Power Systems	
Instruments and Measurements	Mining and Metal Industry	
Magnetic Amplifiers	Petroleum Industry	
Metallic Rectifiers		Education
Nucleonics		Safety
Solid State Devices		Research

Each part has been indexed separately in the back of that particular part. The three parts are not cross-referenced; hence the user should determine first whether the subject matter of the paper desired is in the field of communication and electronics, applications and industry, or power apparatus and systems. Papers are listed in the subject index under several key words in the titles. The original numbers assigned to the papers are given in the author index. Volume 74 contains the technical papers and related discussions presented at these meetings:

1. Winter General Meeting, New York, N. Y., January 31–February 4, 1955.

2. Southern District Meeting, St. Petersburg, Fla., April 13–15, 1955.

3. Middle Eastern District Meeting, Columbus Ohio, May 4–6, 1955.

4. Summer General Meeting, Swampscott, Mass., June 27–July 1, 1955.

5. Pacific General Meeting, Butte, Mont., August 15–17, 1955.

6. Fall General Meeting, Chicago, Ill., October 3–7, 1955.

Statements and opinions given in the papers and discussions published in *Transactions* are the expressions of the contributors for which the American Institute of Electrical Engineers assumes no responsibility.

Analysis of Backlash in Feedback Control Systems With One Degree of Freedom

L. M. VALLESE
ASSOCIATE MEMBER AIEE

Synopsis: The sinusoidal analysis of feedback control systems with one degree of freedom, involving backlash, is developed in the time instead of in the frequency domain, by application of the Kryloff-Bogoliuboff method.[10] Some interesting relations among the parameters of the output response are found, and the conditions for the limit of stability are derived. An approximate method to take into account the harmonic content of the input is indicated.

NONLINEARITIES which occur in feedback control systems can be classified broadly as single valued and multivalued, according to the nature of their input-output relationship. Saturation and backlash are typical respective examples of these two classes. While for single-valued nonlinearities such a relationship is well defined and expressed only in terms of the instantaneous value of the input, the same is not true for multivalued nonlinearities. In this case the relationship depends both upon the instantaneous value and the rate of change of the input, i.e., upon its wave form, and in general cannot be expressed in simple analytical form. As a consequence, many of the iteration and perturbation procedures of nonlinear analysis, which are used successfully in the case of single-valued nonlinearities, cannot be applied in problems involving multivalued nonlinearities.

The usual method of attack of such problems is based on the assumption that the input quantity varies monotonically with time between two equal and opposite peak values; then, the input-output relationship assumes well-defined and simple expression and graphical representation, depending only upon the magnitude of the peak value of the input quantity. Step-by-step analyses of the autonomous behavior of feedback control systems with one degree of freedom, involving backlash in the forward or in the feedback branch of the loop, have been described in recent literature.[1-4] The validity of these rigorous methods is subject to the correctness of the assumed backlash characteristic.

The specification of monotonic variation can be satisfied, in particular, with the assumption that the input quantity varies sinusoidally with time. In this case it is possible to apply an approximate analysis, based on the principle of harmonic balance and subject to the before-mentioned limitation. Such a procedure, developed for feedback control systems by Tustin,[5] Godfarb,[6] Kochenburger,[7] and Johnson,[8] has proved very useful and sufficiently accurate to replace the more cumbersome step-by-step methods. By introducing the concept of describing function, which can be considered as the transfer function corresponding to the fundamental Fourier component of the output, the procedures of frequency analysis can be extended to nonlinear systems, providing in general graphical solutions for the amplitude and the frequency of the input quantity.

The aim of this paper is to develop in the time instead of in the frequency domain the sinusoidal analysis of a feedback control system with one degree of freedom, containing backlash. This procedure, which was first proposed by Van der Pol[9] and modified by Kryloff and Bogoliuboff,[10] not only provides analytical instead of graphical solutions but also permits the harmonic content of the input quantity on the multivalued characteristic to be taken into consideration up to a certain extent.

For purposes of exemplification it will be assumed that the backlash nonlinearity is in the feedback branch of the loop; see Fig. 1. The differential equation representing the behavior of the system is of type

$$\ddot{x} + a_1\dot{x} + a_2 y = a_2 x_i(t) \qquad (1)$$

where $y = y(x)$ is the functional relationship pertinent to the nonlinear term, and $x_i(t)$ is the forcing term which is to be equated to zero when studying the autonomous behavior.

Discussion

In Fig. 2 the characteristic of backlash associated with an input wave form $x = A_x \sin \Omega t$ is indicated. It is symmetric with respect to the origin and, if $2x_0$ is the free play, possesses intercepts x_0 and y_0 with the co-ordinate axes, which are fixed parameters of the system. In addition $y_0 = Kx_0$, where $K = \tan \alpha$.

For the following analysis it is convenient, in general, to represent a multivalued nonlinearity $y = y(x)$ as the sum of a single-valued mean relationship and a multivalued alternating relationship, where either or both can be nonlinear. For example, the single-valued characteristic associated with backlash is $y = Kx$, and the multivalued alternating nonlinear characteristic, according to

$$y = Kx + f(x) \qquad (2)$$

is for sinusoidal input

$$
\begin{aligned}
f(x) &= -Kx_0, && \text{for } -\psi_1 \leq \psi \leq \pi/2 \\
&= K(A_x - x_0 - x), && \text{for } \pi/2 \leq \psi \leq \pi - \psi_1 \\
&= Kx_0, && \text{for } \pi - \psi_1 \leq \psi \leq 3\pi/2 \\
&= -K(A_x - x_0 + x), && \text{for } 3\pi/2 \leq \psi \leq 2\pi - \psi_1
\end{aligned}
\qquad (3)
$$

The phase angle ψ_1 is defined by

$$\sin \psi_1 = 1 - 2x_0/A_x$$

In Fig. 3 $f(x)$ is plotted against x and against ψ for the case $A_x = 3x_0$, $K = 1$. It is of interest to observe that $f(\psi)$ is periodic and single valued, and possesses a Fourier spectrum of odd harmonics only, i.e.

$$f(\psi) = \sum_{n=1,3,5}^{\infty} (f_n(A_x)\cos n\psi + g_n(A_x)\sin n\psi) \qquad (4)$$

since

$$f(\psi) = -f(\psi + \pi)$$

The separation of $y(x)$ in a mean single-valued part and in an alternating multivalued part is important to define a relative measure of multivalued nonlinearity. The latter is useful for the determination of the error involved in the approximate analysis. In the case of the backlash characteristic one can take as a relative measure the quantity $\int_{-A_x}^{A_x} \dfrac{f(x)}{Kx}\, dx$ which is smaller the larger the ratio A_x/x_0 is.

Substitution of equation 2 in 1 provides

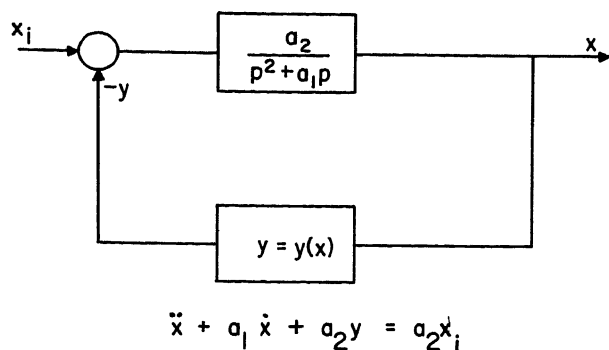$$\ddot{x} + a_2 Kx = -a_1\dot{x} - a_2 f(x)$$

Fig. 1 (above). Feedback control system with one degree of freedom, containing backlash

Fig. 2 (right). Backlash characteristic for sinusoidal inputs of various amplitudes

$$\ddot{x} + a_1 \dot{x} + a_2 y = a_2 x_i$$

where it is assumed that $a_1\dot{x} + a_2 f(x)$ remains small for all values of time. Letting, in accordance with the method of Kryloff and Bogoliuboff

$$x = A_x(t) \sin [\omega t + \varphi(t)] = A_x \sin \psi$$
$$\dot{x} = A_x \omega \cos \psi$$

where A_x and φ are assumed to vary "slowly" with time, one has in first approximation

$$\frac{dA_x}{dt} = \frac{-1}{2\pi\omega} \int_{-\psi_1}^{2\pi-\psi_1} [a_1\omega A_x \cos \psi +$$
$$a_2 f(A_x \sin \psi)] \cos \psi d\psi \quad (5)$$

$$\frac{d\psi}{dt} = \Omega = \omega + \frac{1}{2\pi\omega A_x} \int_{-\psi_1}^{2\pi-\psi_1} [a_1\omega A_x \cos \psi +$$
$$a_2 f(A_x \sin \psi)] \sin \psi d\psi$$

Substituting the relations 3 and performing the integrations

$$\frac{dA_x}{dt} = \frac{1}{2\pi\Omega A_x} [-\Omega a_1 \pi A_x{}^2 + 4Ka_2 x_0(A_x - x_0)] \quad (6)$$

$$\Omega = \omega + \frac{Ka_2}{4\pi\omega} (\sin 2\psi_1 + 2\psi_1 - \pi) \quad (7)$$

where $\omega = \sqrt{Ka_2}$ has been replaced with Ω in equation 6. To investigate the existence of sustained oscillations one has to solve the equation $(dA_x)/(dt) = 0$; letting $\rho = Ka_2/a_1{}^2$ and $\Omega^1 = \Omega/a_1$, there follows
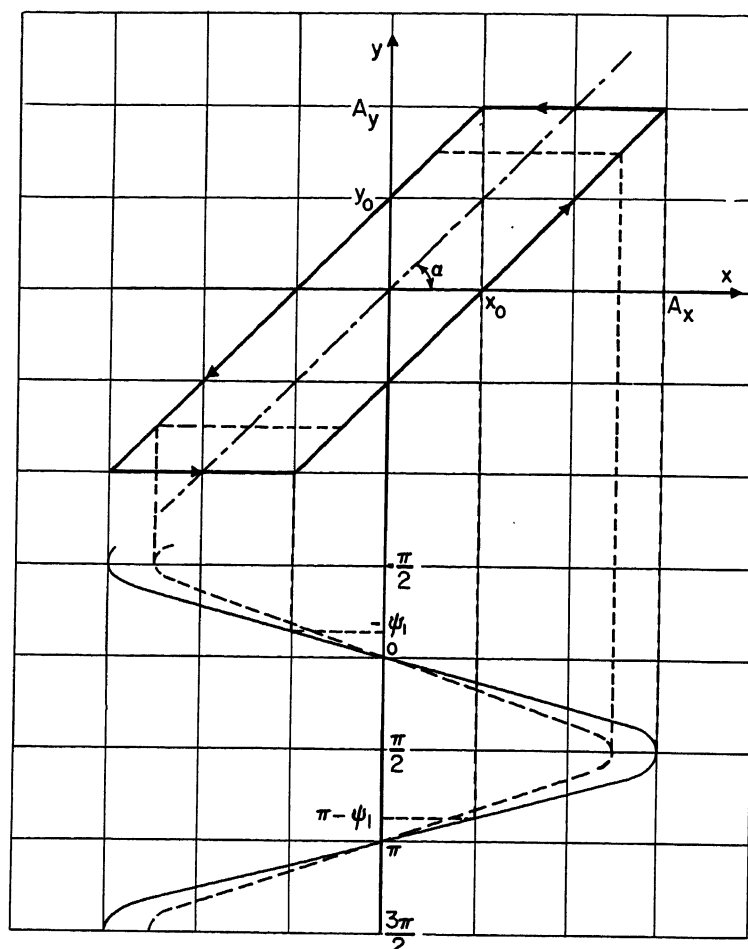
$$\frac{A_x}{x_0} = \frac{2\rho}{\pi\Omega^1} \left(1 \pm \sqrt{1 - \frac{\pi\Omega^1}{\rho}}\right) \quad (8A)$$

$$\Omega^1 = \sqrt{\rho} \left(\frac{3}{4} + \frac{2\psi_1 + \sin 2\psi_1}{4\pi}\right) \quad (8B)$$

From equation 8(A) there follows the condition

$$\pi\Omega^1 \leq \rho$$

In addition it is found that A_x/x_0 is uniquely determined as a function of the adimensional parameter $\xi = \rho/\Omega^1$.

In Fig. 4 the quantities A_x/x_0, ρ, and Ω^1 have been plotted in separate curves as functions of ξ. Each of these curves possesses two branches, i.e., provides two solutions for a given value of ξ. However, from the relation

$$\sin \psi_1 = 1 - \frac{2x_0}{A_x}$$

which combined with equation 8 can be written $\sin \psi_1 = \pm\sqrt{1 - \pi/\xi}$, it is seen that these two solutions correspond to equal and opposite ψ_1 values, positive for the upper branch and negative for the lower branch respectively. For negative ψ_1 values the amplitude A_x is less than the free play $2x_0$; see Fig. 5.

The stability of the found solution can be investigated on the basis of the sign assumed by $[d/(dA_x)] [(dA_x)/(dt)]$ in correspondence with the amplitude A_x. The solution is stable or not depending on whether this sign is negative or positive. From equations 6 and 7 one obtains

$$\frac{d}{dA_x} \left(\frac{dA_x}{dt}\right) = -\frac{a_1}{2} \left[1 - \frac{4\xi}{\pi} \left(\frac{x_0}{A_x}\right)^2 + \frac{4}{\pi^2} \left(1 - \frac{x_0}{A_x}\right) \frac{\xi^2(1 + \cos 2\psi_1)}{A_x{}^2 \cos \psi_1} x_0{}^2\right] \quad (9)$$

This expression becomes zero for $A_x =$

$1.6x_0$ (which corresponds to the point of minimum of the ρ curve in Fig. 4, i.e., $\xi \cong 1.05$, $\rho \cong 5.0$), and is $\lessgtr 0$ respectively for $A_x \gtrless 1.6x_0$. Therefore, according to the results of this analysis, for values of $\rho \gg 5$ the system can settle asymptotically on either of two oscillations, one unstable with amplitude less than $1.6x_0$ and one stable with larger amplitude. Physically, of course, the unstable oscillation is not observed. No steady-state oscillations are found for $\rho < 5$. As an example, the computation of A_x/x_0 and Ω^1 is shown in Fig. 4 for a given value of $\rho = 20$.

For design purposes the accurate evaluation of the minimum value of ρ for existence of oscillations is very important. Unfortunately the present analysis is not well suited to this task since, at this limit, the relative measure of multivaluedness is very large. In Fig. 4 the results found with step-by-step analysis[4] have been indicated with dashed curves. It is seen that complete agreement is obtained only for $A_x \gg 2x_0$, that the unstable oscillation actually does not exist, and that the critical value of ρ is ~3.046.

Aside from the limitations indicated, the method of Kryloff and Bogoliuboff
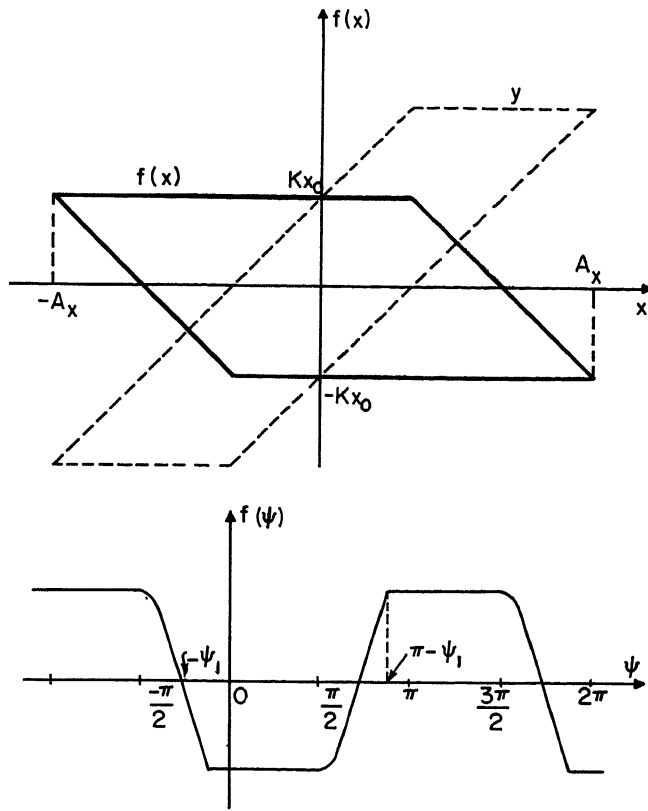
Fig. 3. Plot of the multivalued function f(x) against x and against ψ

Fig. 4. Plot of A_x/x_0, ρ, and Ω^1 versus $\dfrac{A_x}{x_0}=\dfrac{2\xi}{\pi}\,[1\pm\sqrt{1-(\pi/\xi)}]$.

$$\Omega^1=\rho/\xi,\ \rho=\xi^2\left(\frac{3}{4}+\frac{2\psi_1+\sin^2\psi_1}{4\pi}\right)^2$$

possesses definite advantages of elegance of treatment. In addition, for $\rho\gg5$, it provides a procedure for an approximate correction of the basic assumption of sinusoidal input made in connection with the backlash characteristic. As a matter of fact, from the theory it is found that a refinement of the solution of first approximation

$$x=A_x\sin\psi$$

is given by the expression

$$x=A_x\sin\psi+\sum_{n=3,5...}^{\infty}\frac{1}{(n^2-1)\Omega^2}\ (f_n(A_x)\times$$
$$\cos n\psi+g_n(A_x)\sin n\psi)\quad(10)$$

where $f_n(A_x)$ and $g_n(A_x)$ have been defined in equation 4, and A_x and Ω are the values determined in equation 8. For example, one has

$$f_3(A_x)=\frac{1}{\pi}\int_{-\psi_1}^{2\pi-\psi_1}f(A_x\sin\psi)\cos 3\psi\,d\psi$$
$$=\frac{KA_x}{\pi}\left(-\frac{4}{3}\sin 3\psi_1\,\frac{x_0}{A_x}+\frac{1}{12}\times\right.$$
$$(-1+8\sin 3\psi_1-6\cos 2\psi_1+3\cos 4\psi_1)$$

$$g_3(A_x)=\frac{1}{\pi}\int_{-\psi_1}^{2\pi-\psi_1}f(A_x\sin\psi)\sin 3\psi\,d\psi$$
$$=\frac{KA_x}{\pi}\left(-\frac{4}{3}\cos 3\psi_1\,\frac{x_0}{A_x}+\frac{1}{12}\times\right.$$
$$(8\cos 3\psi_1+6\sin 2\psi_1-3\sin 4\psi_1)$$

On the basis of equation 10 the backlash characteristic of Fig. 2 and its representative equations 2 and 3 can be re-evaluated obtaining a new representation of type
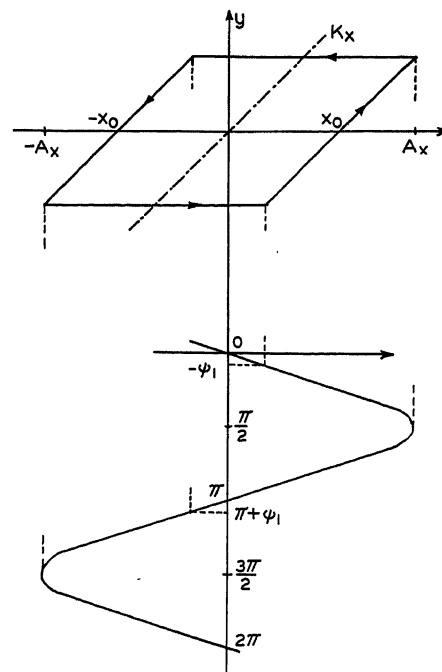
$$y=Kx+f^*(A_x,\ \psi)\quad(11)$$
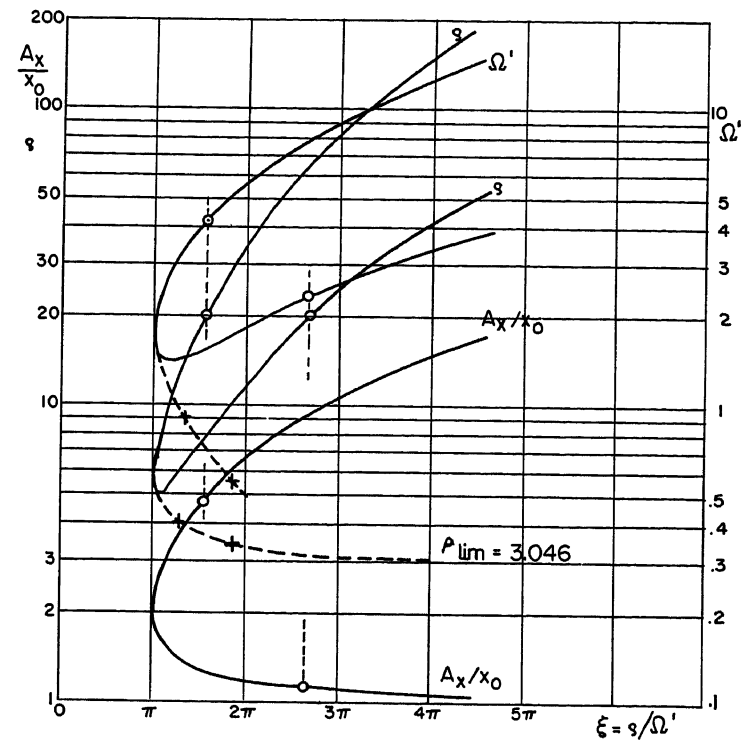

Fig. 5. Backlash characteristic for $A_x<2x_0$

In equation 11 $f^*(A_x,\ \psi)$ can be expanded in a Fourier series containing only odd harmonics. After this has been obtained, equation 1 can be solved by application of the principle of harmonic balance. Letting

$$x=A_x'\sin\psi+\sum_{n=3,5...}^{\infty}\cdot(B_n\cos n\psi+C_n\sin n\psi)$$

where A_x', B_n, and C_n are considered unknown, and substituting into equation 1 one equates to zero the coefficients of $\sin n\psi$, $\cos n\psi$, $(n=1,3,5...)$ and obtains a system of equations in finite terms, which provide as solutions corrected values of A_x, Ω, B_n, and C_n.

Conclusion

The method of Kryloff and Bogoliuboff and that of harmonic balance have been used to solve approximately problems involving nonlinear differential equations with multivalued terms. As an example, the problem of backlash in feedback control systems has been solved. The method is limited to systems with one degree of freedom. It provides the solutions in analytical form, and for this reason some fundamental relationships existing among the various parameters can be pointed out.

References

1. DYNAMICS OF AUTOMATIC CONTROL (book), R. C. Oldenbourg, H. Sartorius. American Society of Mechanical Engineers, New York, N. Y., 1948, pp. 168–73.

2. INSTRUMENT INACCURACIES IN FEED-BACK CONTROL SYSTEMS WITH PARTICULAR REFERENCE TO BACKLASH, H. Tyler Marcy, Morris Yachter, Jerome Zauderer. *AIEE Transactions*, vol. 68, pt. I, 1949, pp. 778–88.

3. SYMPOSIUM ON AUTOMATIC AND MANUAL CONTROL. Academic Press Inc., New York, N. Y., 1951, "An Introduction to the Analysis of Nonlinear Closed Cycle Control," W. E. Scott.

4. BACKLASH IN A VELOCITY LAG SERVOMECHANISM, Nathaniel B. Nichols. *AIEE Transactions*, vol. 72, pt. II, 1953 (Jan. 1954 section), pp. 462–67.

5. THE EFFECTS OF BACKLASH AND OF SPEED-DEPENDENT FRICTION ON THE STABILITY OF CLOSED-CYCLE CONTROL SYSTEMS, A. Tustin. *Journal*, Institution of Electrical Engineers, London, England, vol. 94, pt. IIA, 1947, pp. 143–51.

6. ON SOME NON-LINEAR PHENOMENA IN REGULATORY SYSTEMS, L. C. Goldfarb. *Avtomatika i Telemechanika*, Moscow, USSR, vol. 8, no. 5, Sept.-Oct. 1947, pp. 349–83. Also *Report 1691*, National Bureau of Standards, Washington, D. C., May 29, 1952.

7. A FREQUENCY RESPONSE METHOD FOR ANALYZING AND SYNTHESIZING CONTACTOR SERVOMECHANISMS, Ralph J. Kochenburger. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 270–84.

8. SINUSOIDAL ANALYSIS OF FEEDBACK-CONTROL SYSTEMS CONTAINING NONLINEAR ELEMENTS, E. C. Johnson. *AIEE Transactions*, vol. 71, pt. II, July 1952, pp. 169–81.

9. ON A TYPE OF OSCILLATION HYSTERESIS IN A SIMPLE TRIODE GENERATOR, B. Van der Pol, E. V. Appleton. *Philosophical Magazine*, London, England, 1921.

10. INTRODUCTION TO NON-LINEAR MECHANICS (book), N. Minorsky. Princeton University Press, Princeton, N. J., 1949.

Basic Circuitry for Electrically Powered Pipe-Line Pump Stations Under Automatic or Remote Control

M. A. HYDE
MEMBER AIEE

W. A. DERR
MEMBER AIEE

THROUGHOUT the last decade there has been an increasing trend to the employment of automatic sequence control systems in electrically powered pump stations on crude oil and products pipe lines. Not only has sequence control actuated by push buttons been widely employed for attended stations but also the use of sequence techniques has been greatly extended by the automatic operation of booster stations and by the remote operation of stations under supervisory control.[1]

The elements in a sequence control system must be arranged in essentially the same manner whether the starting and stopping signals are given by a local operator pressing a push button, by pilot devices responsive to line pressure or flow conditions, or by interposing relays actuated remotely by supervisory control.

This paper presents basic circuitry for electrically powered pipe-line pumping stations which employ sequence control for the pump units and which are therefore suitable for automatic or remote operation. Considerations involved in the application of supervisory control for the remote operation of such stations are presented. The remote control of

tankage facilities at initial and injection stations and of delivery facilities at terminal stations are not covered in this paper.

Electric Elements

Fig. 1 shows a typical single-line electrical diagram for a simple single-unit pumping station, the piping and valve arrangement for which are shown in Fig. 2. The primary disconnecting switch, lightning arresters, and transformers, usually owned by the electric utility company, are located in an outdoor substation which serves the pump station at an appropriate voltage, ordinarily 2,300 or 4,160 volts, depending on the size of the load. The pipe-line company usually installs a set of main secondary disconnecting switches to provide isolation of the station from the power supply.

The scheme shown employs a main circuit breaker 52 located outside the station building and the hazardous area. (See the Appendix for device designations.) The principal control equipment is installed in a pressurized room of the station. For safety the station ventilation is operable even when the main breaker is open. By connecting the auxiliary power transformer on the supply side of the main breaker, this breaker may be opened without interrupting service to the control room and pump room fans, the battery charger, and such lighting as is desired to be continuously availa-

ble. Station isolating valves and local residential load may be likewise served. All equipment supplied from this so-called "essential" auxiliary bus and located within the station building is explosion-proof.

The remaining auxiliaries are supplied through a segregating breaker 52N. Their control equipment is installed in the pressurized control room along with the main pump control. Breaker 52N is manually closed and arranged for tripping simultaneously with main breaker 52 in an emergency such as might cause the control room atmosphere to become hazardous. For this common tripping and lockout operation one or more emergency push buttons are used, explosion proof where installed in the station building. Automatic tripping may be actuated by a relay reflecting a loss of differential air pressure between the control room and pump room, or by a gas analyzer sampling the control room air. These features are discussed more fully in later paragraphs dealing with auxiliary control and protection.

Some installations employ an emergency engine generator set which is located with its control in the nonhazardous area outside the station building and is arranged for automatic starting on failure of utility power. The essential auxiliary power bus is then supplied through contactors interlocked to prevent paralleling the emergency generator with the normal power supply.

Some installations, instead of serving the "nonessential" auxiliaries through breaker 52N, employ a separate transformer for this purpose, its primary connected to the main station bus on the load side of breaker 52. A disadvantage of this arrangement is that the nonessential auxiliaries are de-energized whenever breaker 52 is opened. Except for an emergency condition involving hazardous atmosphere in the control room, it is advantageous for main breaker 52 to open without affecting the auxiliaries. For

example, ordinarily when 52 opens, shutting down the main pump unit (or units), it is desirable to have power available for closing the suction and discharge valves.

The main breaker feeds the pump unit (or pump unit bus) and typically is equipped with relaying for phase reversal and voltage failure 47, overcurrent 51S, and with a voltmeter V and a wattmeter W to indicate power supply and load conditions. An operating transformer on the supply side of the circuit breaker furnishes power for breaker closing, and a battery furnishes direct current for breaker tripping. The main breaker has connected to its load side the main pump motor starting equipment, with a surge protection capacitor and lightning arrester. The single-line diagram, Fig. 1, indicates across-the-line starting of the pump motor although reduced voltage starting methods are often employed.

In a single-unit station employing full-voltage starting of the main unit, the layout shown in Fig. 1 may be modified by combining the main breaker function with the motor breaker, thereby eliminating a breaker. Some installations, particularly new stations designed for unattended operation, have all the control equipment installed in a single location sufficiently separated from the pumping equipment proper to be outside of the hazardous area. With this arrangement, no hazard is introduced in the control room when its ventilation is interrupted, and the segregating breaker 52N is omitted. Other local differences in individual projects introduce deviations in detail from the layouts discussed, but they are generally representative of pipe-line practice.

Auxiliaries and Their Control

As is already evident, the auxiliaries and the control schemes employed for these devices vary according to the station layout. Some typical arrangements commonly used for the principal auxiliaries will be described.

PUMP ROOM VENTILATION

Treatment of pump room exhaust fans serving explosion-proof motor installations is discussed later as a part of the pump unit starting sequence. Installations employing force-ventilated motors in the pump room require separate ventilation of the pump room, which should be preferably by a pair of exhausters with automatic discharge louvers, one unit serving as a spare. Frequently the motors are 2-speed to accommodate seasonal temperature conditions. They should be controlled from maintained-contact switches to insure resumption of operation after a voltage dip or power outage.

CONTROL ROOM VENTILATION

For installations typified by Fig. 1, where the control room is in the same building with the pumps, the control room air should be held at a small but positive pressure differential above the pump room, which may be as low as 1/8 inch water, to prevent any infiltration of hazardous atmosphere from the pump room. The installation should include means of protection against failure of control room ventilation. This may be in the form of a low-pressure differential relay, which actuates an alarm in an attended station, and shuts down an unattended station. This device operates through a time-delay auxiliary relay to permit a temporary loss of differential incident to the normal use of doors. The fan capacity should be adequate for warm weather conditions with the control room closed, and the excess air above exfiltration allowed to escape through automatic louvers. Installation of duplicate fan units is desirable. Control should be by maintained contact switch to provide automatic restarting upon power resumption after an outage.
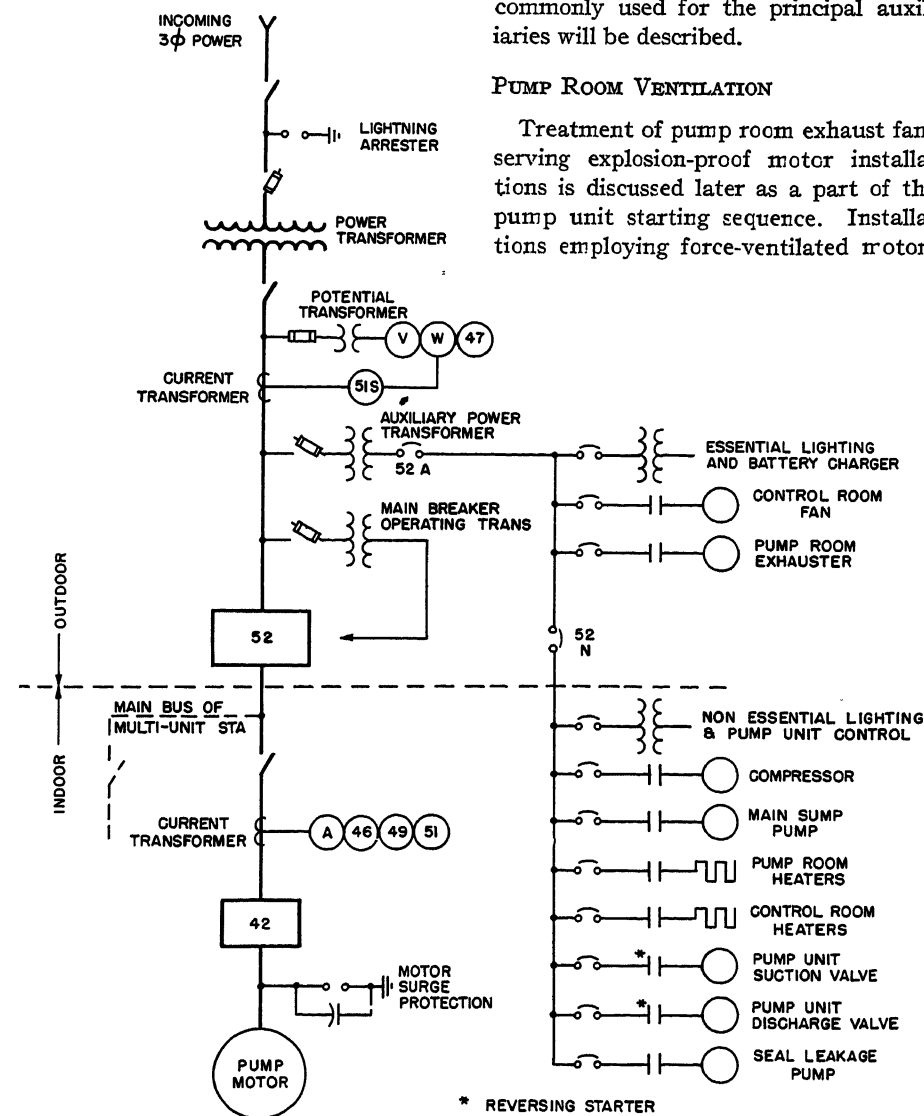

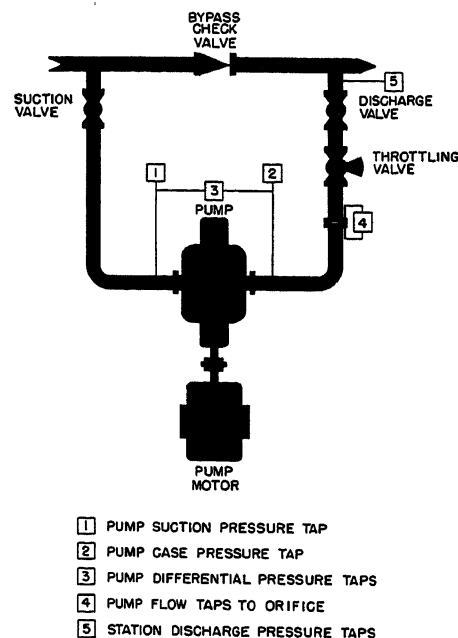Fig. 1. Typical single-line electrical diagram for a simple single-unit pump station

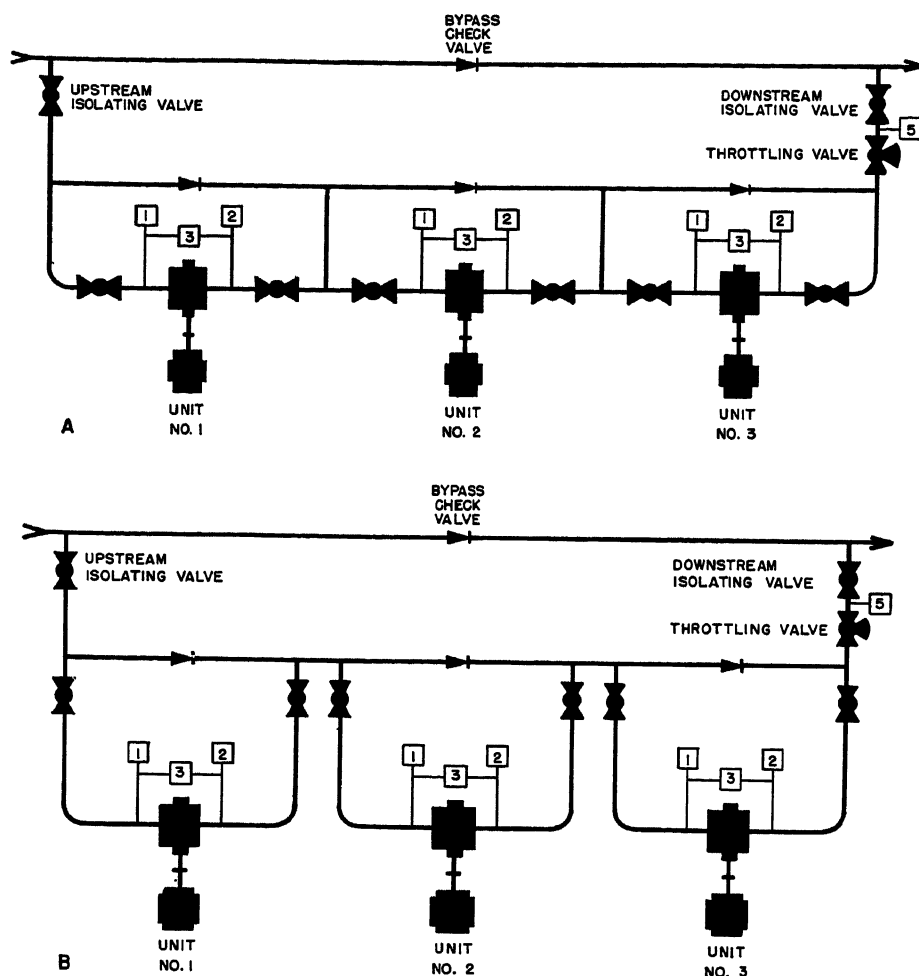Fig. 2. Piping and valve arrangement for simple single-unit pump station

Fig. 3. Typical piping diagrams for multiunit pump stations

In some installations a combustible gas alarm is used to warn the attendant or to shut down an unattended station upon the incidence of a hazardous atmospheric condition. Circuitry for incorporation of these protective devices into the station protective system is detailed later in this paper.

INSTRUMENT AIR COMPRESSOR

Pneumatic control systems are prevalent for the regulation of station pressures. The necessary air compressor, including sometimes a stand-by unit, is controlled automatically by a pressure switch on the air receiver. Air filtering and dehydrating equipment requires servicing, which in unattended stations can be handled on a periodic visit schedule. Instrument air failure should be alarmed by a separate pressure switch, and in the case of unattended stations, shutdown should result.

MAIN SUMP SYSTEM

A level in the main station sump above the normal working maximum should be alarmed by a separate float switch. The problem of sump pump control varies with

different installations. On product lines, and on some crude lines, the possibility of contamination due to batch change in the line generally precludes automatic starting of the sump pump. One solution is to alarm at an upper working level, after which in an attended station the operator can start the sump pump at a suitable time depending upon the batch status. The pump control can be arranged for automatic stopping responsive to a low-level contact. In unattended stations, main sump pumping may be done by local control on the occasion of

inspection visits or, if desired, the sump pump can be started and stopped by supervisory control with supplemental automatic stopping by low-level switch. For this latter type of operation it may be desirable that the controlling location have telemetered information characterizing the line content so that contamination can be prevented.

SEAL LEAKAGE SUMP SYSTEM

For detection of excessive seal leakage it is common to drain the seals of each main pump to a small individual sump. This seal leakage sump may in turn drain to the main station sump through a flow-limiting orifice which will pass the normal leakage rate but will cause a rise in level sufficient to operate a float switch if the leakage is excessive. The float switch in turn effects an alarm and shutdown.

In some installations, to minimize contamination the seal drainage is repumped directly from the small seal sump into the suction side of the main pump by a displacement pump driven by a fractional-horsepower motor. The repump has capacity for ordinary rates of seal leakage, but is inadequate to handle the excessive leakage incident to seal failure, and the resulting rise in sump level above the normal working maximum causes a float switch to alarm and shut down the unit. Some installations provide seal failure protection by a pressure switch responsive to pressure in the housing surrounding the outer end of each seal.

STATION ISOLATION VALVES

It is common to employ "header gate" valves to provide for isolation of the station from the line, particularly at multiple-unit stations. When motor operated, as is particularly desirable in large sizes, these valves may be controlled by a push button at the valves, and attended stations often provide supplementary control from the station control room. For unattended stations, remote super-
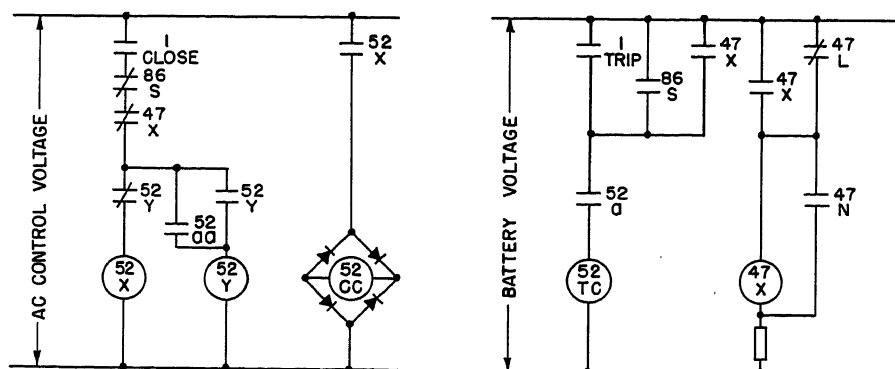

Fig. 4. Schematic diagram for control of main circuit breaker of pump station
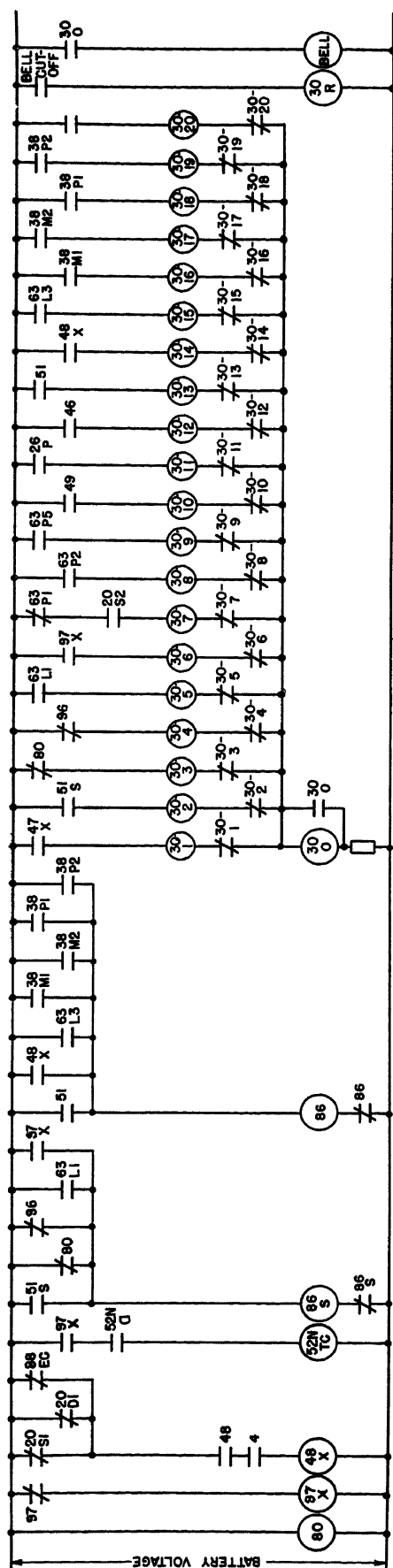
visory control has the advantage of providing rapid isolation of the station in emergency. Individual stations vary as to arrangement of these valves. A common arrangement is one upstream and one downstream of a station by-pass check valve, as shown in Figs. 3(A) and (B). Occasionally a motor-operated by-pass valve is used instead of a by-pass check valve, in which case interlocking provides against the by-pass valve being closed except when both isolating valves are fully open, and against either isolating valve being closed except when the by-pass valve is fully open.

Main Breaker Control

The main circuit breaker may be manually controlled either locally or remotely. Fig. 4 shows a scheme, in which momentary contact 1-CLOSE of master control device 1 energizes the breaker-closing relay $52X$ through contacts of station lockout relay $86S$ and a phase reversal and undervoltage auxiliary relay $47X$. If the breaker is subsequently tripped by any protective relay, it must be manually reclosed. (The protective shutdown system will be described later). A condition for closing the breaker is that power supply voltage and phase sequence be normal; this is detected by device 47, which has a contact $47L$ closed on low voltage or phase reversal, and a contact $47N$ closed when these conditions are normal. These contacts control the auxiliary relay $47X$, as shown in Fig. 4, to prevent closure of the breaker under abnormal supply voltage.

The scheme shown in Fig. 4 can be readily modified for automatic operation of the breaker after it has been closed by replacing momentary contact 1-CLOSE with a contact which remains in the closed position after operation. In this case, unless the breaker has been locked out by device $86S$, it will automatically reclose when voltage and phase sequence are normal.

Pump Unit Control

Fig. 5(A) shows a typical scheme for automatic starting and stopping of a

Fig. 6. Schematic diagram for station and unit lockout and annunciation

single-unit station as shown in Fig. 2, with sequence operation of the unit suction and discharge valves and a solenoid-operated valve for venting the pump, as well as an associated pump room exhaust fan. Whether operation is entirely manual or by automatic sequence, the operating procedures should be carried out in prescribed order.

For starting:

1. Open suction valve and pump venting valve, and start necessary ventilation equipment.
2. Energize pump motor.
3. Open discharge valve and close pump venting valve.

For stopping:

1. De-energize pump motor and its associated ventilation equipment.
2. Close discharge valve.
3. Close suction valve.

There are individual variations from these procedures; e.g., many installations do not require venting, and others in stopping the unit close the suction and discharge valves simultaneously. Fig. 5(A) is based on the sequences just given, and shows a circuit breaker as the switching device 42 in a full voltage starting system. With appropriate minor modifications the diagram can be made to apply to control apparatus using contactor-fuse combinations for motor starting, and to various methods of reduced-voltage starting.

Fig. 5(A) applies to an installation where ventilation is required to remove the motor loss heat, as for an explosion-proof motor located in the pump room. In this case, an air intake should be provided in the wall opposite the motor, and an exhauster driven by an explosion-proof motor should be installed in the wall opposite the pump. Such provision is of course unnecessary for motors located outdoors.

STARTING SEQUENCE

The starting signal originates with momentary closure of master element 1-START contact, which can be a contact of a local push button, an automatic pilot device, or a remotely actuated relay. In series with this contact is limit switch contact 20D4 to prevent starting the unit when the discharge valve is in other than the closed position. This is desirable to reduce the starting load on the motor. When the 1-START contact closes the circuit through 20D4, master relay 4 seals itself in, starts the unit exhaust fan by energizing its contactor 88EC, and prepares a circuit to energize the OPEN contactor 20S0 of the suction valve controller.

This circuit is completed through an auxiliary switch on the exhaust fan contactor 88EC. If desired, a vane switch or auxiliary contact on the exhaust damper may be used instead of a contact of the exhaust fan contactor 88EC. At the end of the opening stroke 20S0 is de-energized by the opening of limit switch 20S1 on the valve mechanism. Closure of relay 4 sets up a circuit for energizing solenoid vent valve 20V, and as the suction valve leaves its closed position the closure of limit switch 20S3 energizes the vent valve solenoid to open the vent valve. The vent valve solenoid is subsequently de-energized, closing the vent valve, by the opening of limit switch 20D4 when the discharge valve leaves the closed position. Closure of relay 4 prepared a circuit for energizing relay 42X of the pump motor breaker, which circuit is completed through limit switch contact 20S2 when the suction valve reaches the open position. Closure of relay 42X causes the breaker to close.

When the breaker closes, an auxiliary switch 42a closes to energize the OPEN contactor 20D0 of the discharge valve controller, and at the end of the opening stroke 20D0 is de-energized by the opening of limit switch 20D1 on the valve mechanism. The pump unit is then in operation.

A sequence completion check is initiated when master relay 4 is energized, completing a circuit to timing relay 48, which is set to provide a sufficient interval for normal completion of the starting sequence. If, at the termination of this interval, the suction and discharge valves are not in the full open position and the exhaust fan contactor is not closed, the unit is shut down through lockout relay 86. See Fig. 6 and later discussion of protection.

When the pump motor is force-ventilated, it is equipped with a motor-driven fan located outside the hazardous area. The unit control sequence then must provide for purging the duct system and motor housing for a predetermined time before the pump motor can be started; during this interval the suction valve is opening. For this arrangement, the circuits controlling motor breaker 42 are modified as shown in Fig. 5(B). Closure of master relay 4 energizes contactor 88BC which starts the force-ventilating fan. When air flow is established through the duct system, flow switch 88FF2 closes, energizing timer 88FX which, after an interval sufficient to purge the duct system, closes to energize breaker-closing relay 42X, the suction valve having meanwhile reached the open position. Upon

completion of the starting sequence, the unit is on the line and the pump operation is subject to the hydraulic system control which will be discussed later.

STOPPING SEQUENCE

The stopping sequence can be initiated by momentary closure of master element 1-STOP contact, by opening of the main breaker 52, by operation of lockout relay 86, or by any of the other protective devices shown connected across the coil of master relay 4 in Fig. 5(A). When relay 4 is released by closure of any of the contacts shown, the trip coil 42TC of the motor breaker is immediately energized and the exhaust fan contactor 88EC is released, sequence checking relay 48 reset, and a circuit is prepared for energizing the CLOSE contactor 20DC of the discharge valve. This circuit is completed by auxiliary switch 42b when the motor breaker opens. When the discharge valve reaches the closed position it is stopped by opening of limit 20D3, at which time 20D4 completes a circuit to the CLOSE contactor 20SC of the suction valve. Closure of the suction valve completes the stopping sequence. After initiation of a stopping signal the unit cannot be restarted until the discharge valve is completely closed. The suction valve cannot be opened or closed unless the discharge valve is fully closed. In some installations, as already mentioned, the stopping sequence is modified to provide simultaneous closure of the suction and discharge valves. In some cases the valve employs a torque-actuated limit switch to stop travel in the closed position, and sometimes a similar switch is used to stop travel in the open position.

SCRAPER PASSAGE CONTROL

In some stations automatic stopping and restarting is arranged to accomodate the passage of a scraper. This is accomplished by recently deviced scraper detectors, one of which, installed just upstream of the station, actuates a shutdown relay as the scraper enters the station zone. With the pump down the scraper proceeds through a full-opening station by-pass check valve, and then through a second detector on the outgoing side of the station which energizes a starting relay to restore the station to normal operation.

Co-ordination with Hydraulic Control

With constant-speed motors, which are ordinarily employed for reasons of simplicity and adaptability to hazardous atmosphere, regulation of the station incre-

mental pressure to suit line requirements is usually accomplished by some form of throttling valve in the station discharge. This throttling valve is under pneumatic control by pressure-regulating elements responsive to the station suction and discharge pressures and so arranged as to keep these pressures within prescribed limits under varying conditions in the line. This system may also serve very usefully to limit the rate of release of the pressure increment of an oncoming unit into the line, not only to prevent surges in the line but also to limit the load on the motor during the period when the line pressure is building up or "packing" following the opening of the unit discharge valve. In an attended station this is usually accomplished by the operator gradually increasing the setting of the discharge pressure controller by small increments at intervals of several minutes, until the line has stabilized at the higher pressure level.

It is possible, of course, to perform this operation remotely through a motorized drive on the regulator control-point setting mechanism. On one products line employing single-unit stations, the throttling valve controller is equipped with a geared motor drive which, upon starting the pump unit, is energized to advance the controller setting slowly from an initially low value to the desired maximum, and upon shutdown to return the setting to its initial value. This operation is performed locally, without any remote control or supervision, its progress being observed by telemetered values of the station discharge pressure.

Another practice in single-unit stations is to apply to the hydraulic control system an overriding limiting controller responsive either to flow or to motor load. This arrangement permits the settings of the suction and discharge pressure limit controllers to remain fixed at the station's maximum operating condition. In this system a spring-closed, air-to-open control valve is used, and a solenoid valve so installed that, when energized, it will release the air from the control valve, thereby holding it closed. By energizing the solenoid valve when the suction valve begins to open and until the discharge valve is fully open, the throttling valve is held closed, from which position it can initiate its controlling action most effectively in response to the flow-limiting or load-limiting controller.

Features of Multiunit Stations

Use of two or more pump units introduces some additional considerations.

Multiunit stations usually employ pumps connected in series with a single throttling valve on the downstream side of the final pump, as shown in Figs. 3(A) and (B). The arrangement shown in Fig. 3(A) is commonly used in crude oil lines to minimize station piping and pressure drop, while that shown in Fig. 3(B) is usually employed in products lines (and occasionally crude oil lines) where it is important to minimize contamination. The latter arrangement also permits the location of the unit valves outside the pump room.

As previously stated, a single-unit station may not have a main circuit breaker, but in a multiunit station it is always desirable to have a main breaker. The motor surge protection equipment may consist of a single capacitor–lightning-arrester combination connected to the main bus. Time-delay undervoltage relays may be applied to the individual units with progressive time settings so that following a voltage dip, as many units can be retained on the bus as practical, consistent with reasonable current demand.

In multiunit stations shutdown in response to abnormal pressures is desirably sequential in order to minimize hydraulic shock and to keep as much of the station in operation as practical when shutting down part of the units would result in correction of the abnormal pressure. Shutdown due to low suction or high case pressure in the individual pumps is accomplished in the same manner as for a single-unit station, by pressure switches installed on each side of each unit, at tap points 1 and 2 shown in Figs. 3(A) and (B). Since low station suction pressure is reflected on the suction side of the upstream operating unit, the arrangement just mentioned provides inherently sequential shutdown in the event of low station suction pressure. High case pressure will be sensed initially on the downstream unit, which will result in sequential shutdown in the reverse order to that just described for low station suction pressure. Shutdown due to high station discharge pressure must be actuated by a single pressure switch tapped to the downstream side of the station throttling valve (tap point 5 of Fig. 3). In many stations this switch shuts down all units simultaneously. However, sequential shutdown on high station discharge pressure can be accomplished by the scheme shown in Fig. 5(C).

In this system, the protective device contact 63P5 for high station discharge pressure energizes an auxiliary relay 63P5X which closes the tripping circuits

of the operating units sequentially proceeding downstream. The unit sequence is automatically established by pump differential pressure relays *63P3-1, 63P3-2, 63P3-3*, etc., for the respective units, each of which holds open the circuit to the trip coil of the succeeding downstream unit as long as it is actuated by differential pressure across its own pump. If desired, this same shutdown sequence may also be initiated by a pressure switch tapped to the upstream side of the station throttling valve, with its contact in parallel with pressure switch *63P5*.

For stations employing two or more pump units somewhat more extensive provisions are needed than for single-unit stations to control the operating during line packing. A suggested method is to employ an overriding load controller for each pump unit in order to initiate throttling on excessive load, reflecting high flow, in any unit. During starting of the first unit only, the throttling valve is held closed until the suction and discharge valves are fully open. Another method, employed by a large products line, uses one load controller for a 2-unit station, and the controller is pneumatically adjusted to the appropriate setting for single-unit or 2-unit operation by a solenoid valve actuated from a limit switch at the open position of the discharge valve of the second unit.

Another problem introduced in multi-unit stations is that of automatically accommodating the passage of a scraper. As previously discussed, a single-unit station with a full-opening by-pass check valve can be shut down and restarted by scraper detectors signalling the scraper's entry to and departure from the station zone. However, in a multiple-pump station hydraulic and power-demand considerations make it undesirable to restart units simultaneously. By adding a motor operated by-pass valve between the station discharge and station suction, and arranging for the incoming scraper detector to open this valve and for the outgoing detector to close it, it is possible to leave all the pumps running during the scraper's passage.

Protection

In all types of electric equipment installations protection against abnormal operation or equipment failure is important, but in no type of installation is an adequate protective system more important than in a pipe-line pump station, because of the hazardous nature of the medium transported. This is particularly true for automatic or remotely operated stations where usually no personnel are present.

The protection for individual stations will vary according to the basic arrangement. There are certain protective functions which should be provided in all installations; described here are those considered generally essential for the case of a typical unattended station; it is recognized that individual conditions may warrant deviations from the functions enumerated in the following.

There are four basic operations that may result from the functioning of a protective device:

1. Station shutdown consists of opening the main circuit breaker, which is subject to subsequent reclosure.
2. Station lockout consists of opening and locking out the main circuit breaker, after which the breaker can be reclosed only when the lockout device is manually reset.
3. Unit shutdown consists of stopping the unit and closing its suction and discharge valves in the normal manner, subject to subsequent restarting.
4. Unit lockout consists of stopping the unit and closing its suction and discharge valves, after which the unit can be restarted only when the lockout device is manually reset.

In existing installations, the selection of protective functions which will initiate each of these actions has been based on individual engineering analysis. There are no industry-wide standards governing the protection of pipe-line pump stations. Based on experience, protective functions are assigned to the four basic categories which are believed to be best for typical unattended stations.

STATION SHUTDOWN

This should result only from phase reversal or a-c undervoltage. Protection for both of these conditions is usually provided by a single relay shown as device *47* in Fig. 1. As previously discussed and shown in Fig. 4, operation of device *47* trips the main breaker. Auxiliary contacts of this breaker are employed in the master relay circuits of all pump units to effects their normal shutdown. Referring to Fig. 5(A), a contact *52b* of the main breaker auxiliary switch shunts the master relay coil, and release of this relay trips the motor breaker and closes the suction and discharge valves. When voltage conditions are restored to normal the main breaker can be reclosed either manually or automatically, depending on the method of operation employed, and the individual pump units may then be restarted after their suction and discharge valves have closed.

STATION LOCKOUT

The following conditions should result in station lockout:

1. A-c overcurrent *51S*.
2. Battery undervoltage *80*.
3. Low instrument air pressure *96*.
4. High level in main sump *63L1*.
5. Unsafe atmospheric condition in control room *97* or *98*. (Applies only where control room is subject to hazardous atmosphere.)

Referring to Fig. 6, it will be noted that all of these functions result in operation of station lockout relay *86S*, a contact of which trips the main breaker *52*. This results in releasing the master relay *4* of each individual pump unit as described in the preceding section. In the case of the first four functions just listed, the auxiliary power is not interrupted, and release of master relay *4* is followed by tripping of the pump unit breaker and closure of the discharge and suction valves. However, in the case of the fifth function (unsafe atmospheric condition in the control room) closure of device *97* or *98* contact results in tripping of auxiliary segregating breaker *52N* simultaneously with energization of lockout relay *86S*, as shown in Fig. 6. After this lockout operation, reclosure of the main breaker can be accomplished only when relay *86S* has been manually reset. The auxiliary segregating breaker *52N* must be manually closed.

UNIT SHUTDOWN

The following conditions should result in the shutdown of an individual pump unit:

1. Low suction pressure *63P1*.
2. High pump case pressure *63P2*.
3. High discharge pressure *63P5*.
4. Motor overtemperature *49*.
5. Pump case overtemperature *26P*.
6. Motor phase current unbalance *46*.

Occurrence of any of these conditions results in release of master relay *4* and normal shutdown of the pump unit as shown in Fig. 5(A). In the case of low suction pressure protection, note that device *63P1* has its contact in series with a limit switch *20S2*. This is necessary to prevent functioning of the low suction pressure switch until the suction valve is fully open, at which time normal suction pressure should ordinarily exist.

Following such shutdowns, the unit can be restarted by reclosure of the 1-START contact. As previously discussed, the 1-START device can be a local push button, an automatic pilot device contact, or a remotely actuated relay contact.
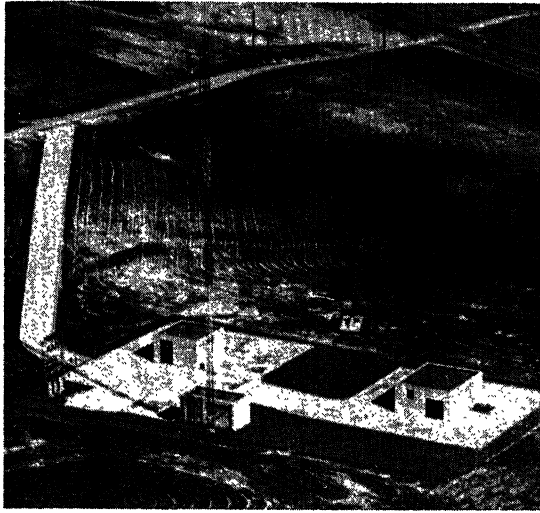
Fig. 7. A single-unit booster station operated by supervisory control over microwave. Note separation of buildings for pump unit, control equipment, and microwave equipment
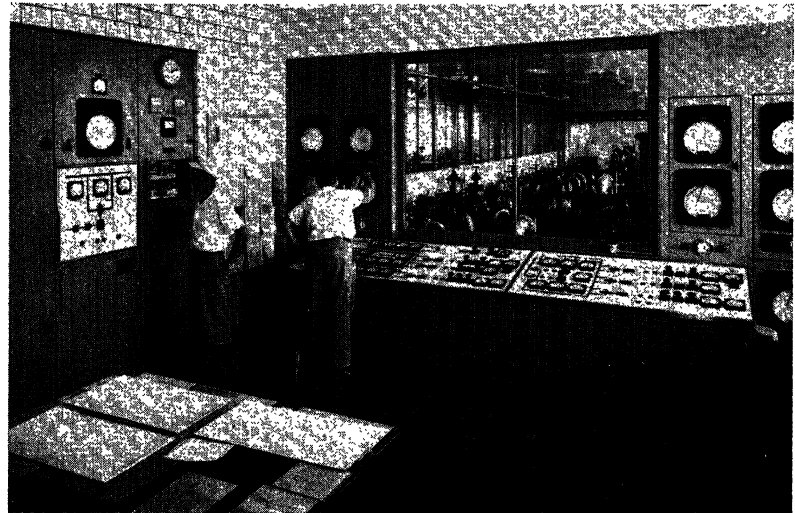

Fig. 8. Control room of a large multiunit pump station. Push-button sequence control for seven local units is provided on central control console. Diagram panel at left controls booster station 56 miles downstream by supervisory control

UNIT LOCKOUT

The following conditions should result in unit lockout:

1. Motor overcurrent *51*.
2. Incomplete starting sequence *48*.
3. Seal failure *63L3*.
4. Motor bearing overtemperature *38M1, 38M2*.
5. Pump bearing overtemperature *38P1, 38P2*.

Referring again to Fig. 6, it will be noted that all of these functions result in operation of pump unit lockout relay *86*, a contact of which shunts the master relay *4*. The release of master relay *4* results in the tripping of the pump unit breaker and closure of the discharge and suction valves. The unit cannot be restarted until lockout relay *86* has been manually reset.

The incomplete sequence protection is initiated by timing relay *48* energized when master relay *4* closes, as shown in Fig. 5(A). After its time setting interval, contact *48* closes in the coil circuit of auxiliary relay *48X*, as shown in Fig. 6. If at this time both the suction and discharge valve limit switches do not indicate the fully open position, or if the contactor *88EC* of the exhaust fan is not in its normal closed position, a circuit is established to energize auxiliary relay *48X* which in turn energizes lockout relay *86*. If during running operation either valve or the exhaust fan contactor leaves its normal position, relay *48X* is energized and lockout results. Where the pump motor is force-ventilated the contact of device *88EC*, in the coil circuit of auxiliary relay *48X*, is replaced by a contact of the air-flow switch *88F*.

A recently developed mechanical vibration-monitoring relay is applicable to initiate lockout upon incipient mechanical failure of a main pump or its driving motor. This device is particularly advantageous for pumps employing ball bearings, since in a ball bearing roughness is usually an earlier indication of trouble than is temperature rise.

LOCAL ANNUNCIATION

Regardless of the method of control, whether local, automatic, or remote, it is desirable that all protective functions be individually annunciated locally. In attended stations this annunciation frequently takes the form of an audible alarm in combination with identifying indicating lights appropriately arranged on a graphic control panel. The controlling location for a remotely operated station is likewise equipped with suitable indicating lights along with the control devices. At the remotely controlled station the most satisfactory form of annunciator is the drop type. All of the protective devices discussed in this paper are shown connected to individual drops of an annunciator in Fig. 6. With the type of annunciator shown, closure of any protective device contact completes a circuit to the coil of an individual drop *30-1*, etc., in series with the operate coil *30-0* of a latching element of the annunciator. When the operate coil is thus energized, contact *30-0* of the latching element closes to increase the current through the individual drop to cause it to operate and open its coil circuit. A second contact *30-0* can be used to sound a bell as shown in Fig. 6 or it can initiate a remote alarm. The alarm circuit is restored to normal by

energization of coil *30-R* which mechanically resets the latching element.

Automatic Operation

As previously indicated, it is possible to have the starting and stopping sequences for a pump unit initiated automatically.[2-4] Elements responsive to line pressure or flow conditions at the pump station can be employed to initiate these sequences, as shown in Fig. 5(A).

Completely automatic starting and stopping is usually only employed in single-unit booster stations. For the usual automatic installation, no information concerning the functioning of the station or the hydraulic conditions which exist is provided at any remote location. The operation of the automatic station is inferred from pressure and flow conditions at the adjacent attended pump stations.

Remote Operation

The unattended remote operation of both single-unit and multiunit pump stations is becoming quite prevalent.[5-8] Such remote operation can be advantageously obtained by the use of supervisory control and telemetering equipment. Supervisory control operating over a single telegraphic-type channel is used to perform the necessary operations, to provide indications of the position of apparatus, and to provide indications and alarms for trouble conditions. Telemetering equipment is used to provide quantitative indications of hydraulic and electrical conditions at the unattended station. The telemetered indications can
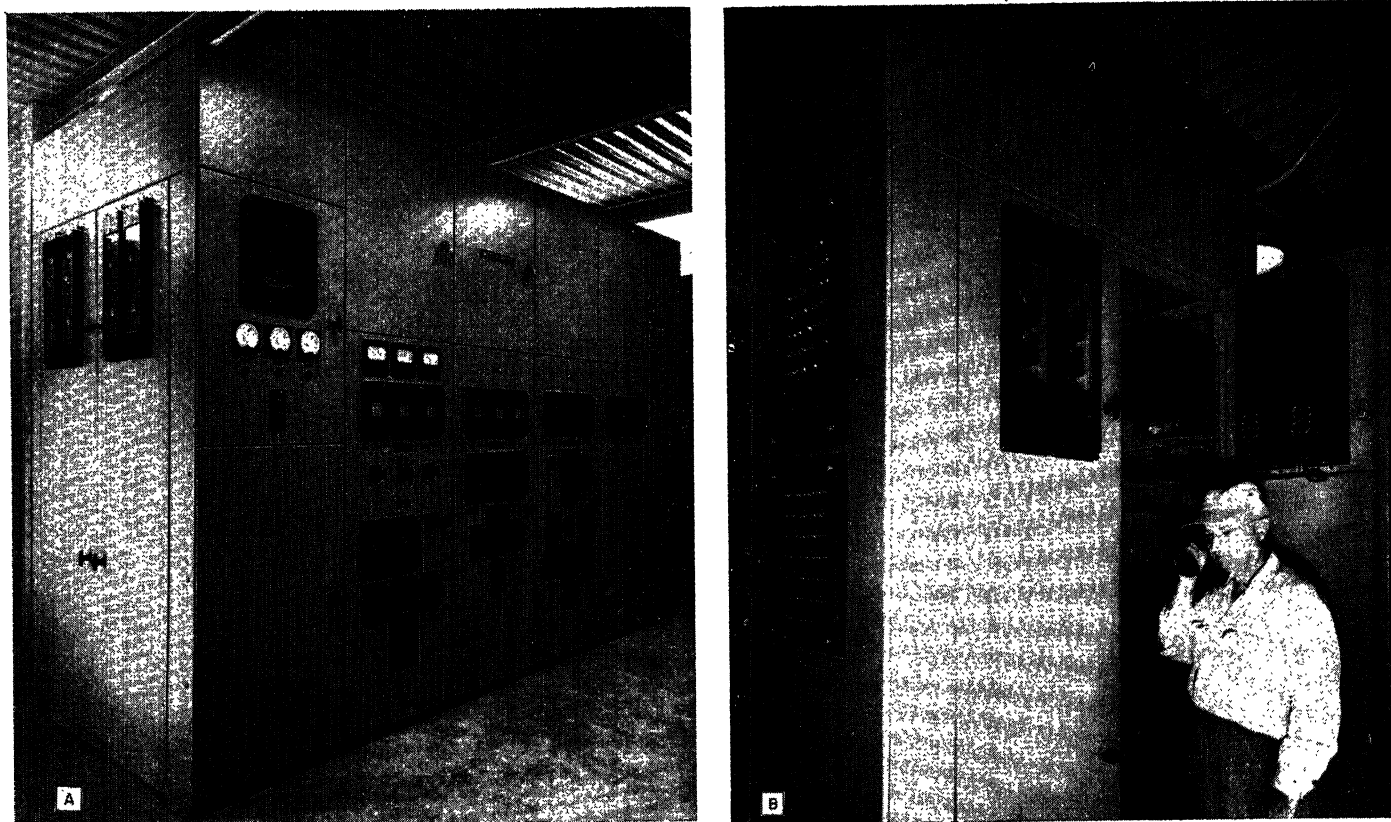
Fig. 9. Control equipment at a remotely operated pump station

A—Switchgear assembly for a single-unit booster station remotely controlled. This assembly combined main station breaker, reduced voltage motor starting equipment, auxiliary power transformer, control center, sequence and protective relays, annunciator, local electric and hydraulic instrumentation, supervisory control, and telemetering transmitters

B—Telemetering transmitters and supervisory relay cases (left) of switchgear assembly shown in (A)

be obtained continuously by the use of a separate channel for each indication, or selectively over the supervisory control channel. In addition, multiplexing techniques can be employed to connect the telemetering transmitters and receivers to the same channel in consecutive order.

In general remote operation of a pump station is not effected from a location closer than an adjacent attended pump station. An increasing trend is the establishment of centralized control locations, such as the initial station or the pipe-line system dispatching office, from which large numbers of attended stations can be operated. Therefore, the remote operation must be effective over relatively long distances ranging from approximately 20 miles to hundreds of miles. The necessary channels for operation of pipe-line pumping stations can be provided by telegraphic-type tones operating over privately owned microwave or telephone line carrier links.[9,10] In addition, leased facilities can be employed in two ways.[11] Individual telegraphic-type channels can be leased or a voice channel can be leased and the necessary number of telegraphic-type channels obtained by the application

of audio-tone generating and receiving equipment.

Fig. 7 shows a view of a single-unit pump station which is remotely operated over a microwave channel. From left to right it shows outdoor substation, control building, microwave building and tower, and pump building.

SUPERVISORY CONTROL

When a pipe-line pumping station is remotely operated by supervisory control, the equipment applied must be arranged to provide the following minimum control functions and indications of device positions:

1. Start-stop sequence control of each pump unit.

2. Indication of stopped or running condition of pump motors.

3. Indication of open, closed, and intermediate positions of the suction and discharge valves of all pump units.

In addition, it is extremely desirable to provide an indication of the position of the main breaker (also control of main breaker unless it is completely automatic) and to provide indications of station lockout and individual pump unit lockout. A super-

visory control cutoff or transfer switch is desirable at the controlled station, and it is usual to indicate position of this switch at the controlling location. There may be installations where it is desirable to control the main sump pump. Unless completely automatic means are provided as previously discussed, it may be necessary to provide for remote adjustment of the control point setting of the station discharge pressure controller. Where electrically operated isolating valves are employed, provisions for their remote operation may be desirable.

Rather than have the starting and stopping of pump units directly under the control of a remote operator, it is sometimes preferable to provide a combined remote manual and automatic arrangement to govern the starting and stopping of pump units. With this type of control the operator can stop a pump unit at any time, but he can only prepare it for starting with the actual starting as well as subsequent stopping then being dependent on line hydraulic conditions as sensed by pressure or flow switches.

Economic considerations usually prohibit individual indications over super-

visory control for all of the protective device operations which are shown in Fig. 5(A) for a typical single-unit station. However, the protective devices can be grouped in various combinations to provide a relatively small number of indications which can assist in determining the type of personnel to be dispatched to the controlled station after lockout is reported. A typical grouping of functions for common indications for a single-unit station is as follows:

1. Power failure: device 47.

2. Electric equipment failure: devices 46, 48, 49, 51S, 51, 80, 88E.

3. Mechanical equipment failure: devices 26P, 38M1, 38M2, 38P1, 38P2, 63L1, 63L3, 96 and 97 or 98.

4. Abnormal hydraulic conditions: devices 63P1, 63P2, 63P5.

Supervisory control for pipe-line pumping is usually arranged to provide the same type of operating procedures as would be required for local push-button sequence control. The equipment is usually arranged so that it is only necessary to operate a start and stop push-button to initiate the starting or stopping sequence for a pump unit. Likewise, the alarm lamps for protective device operation are usually arranged to remain lighted until the trouble contact has opened and a reset push button has been operated. The lamp indications for the pump motors and suction and discharge valves are often located in a diagram of the main station piping. A supervisory control diagram panel for the remote operation of a single-unit downstream pump station is shown at the left of Fig. 8. The control desks shown in Fig. 8 are for the sequence control of local pump units.

The supervisory control equipment is generally installed with the motor control equipment. In the case of new stations, it is often made an integral part of the switchgear assembly. Fig. 9(A) shows the switchgear for a single-unit unattended pump station. The supervisory control equipment is mounted on the rear panel of the left-hand unit of this switchgear. This supervisory control panel is shown at the left of Fig. 9(B).

TELEMETERING

Regardless of the number of pump units in a remotely operated station, it is essential to provide telemetered indications to the controlling location of the following hydraulic quantities:

1. Initial pump suction pressure.

2. Final pump discharge pressure.

3. Outbound line pressure. (Not required if no throttling valve is employed.)

Telemetering of flow may be desirable where flow conditions are not uniform throughout the line as may result from deliveries at intermediate points. While not essential for remote operation of a pump station, there are instances in which it may be desirable to telemeter electrical quantities such as watts and amperes for the pump motors and volts for the station.

As previously indicated, the telemetering may be continuous, the telemetered quantities may be indicated in consecutive order through some type of time-division multiplexing, or they may be selectively obtained at the will of the dispatcher through the supervisory control equipment. Economics usually dictates in favor of the latter two arrangements although there are many possible compromise arrangements involving some continuous and some selective or multiplexed telemetering.

Pulse-duration or frequency-type telemetering systems are usually employed.[12-14] The pulse-duration-type telemetering transmitters may be directly actuated by the hydraulic quantity or by the electric output of a transducer. The frequency-type telemetering transmitters can only be employed for hydraulic quantities through the use of transducers.[15] The frequency-type telemetering system provides faster rate of response but requires channels capable of handling higher keying speeds. Telemetering transmitters of the frequency type are shown on the doors of the switchgear assembly at the left of Fig. 9(A).

Conclusion

Modern electrically powered pipe-line pumping stations are the result of the technical advances in this field which have been made during a period of some three decades. The advent of push-button sequence control for such stations approximately 15 years ago was the forerunner of today's accelerated use of remotely operated unattended pumping stations and completely automatic stations. The justification for such remotely controlled and automatic pumping stations is purely economic.

Modern supervisory control and telemetering systems make it possible to include many desirable features in the design of remotely operated pumping stations regardless of the type of channel employed.

Reliability cannot be overemphasized in the design of remotely operated or automatic stations. A full complement of devices to detect all hydraulic and electrical trouble conditions is all-important. Locally actuated shutdown of the unit or station in trouble should result after a condition of trouble has been detected by one of these protective devices. Remote or automatic operation should only be employed when its functioning is coordinated to an adequate protective system.

Heretofore there has been a lack of consistent practice in designating equipment devices for pipe-line pump stations. As the result of considerable investigation, designations have been employed in this paper based on ASA Standard C37.2[16] with such adaptations as were deemed desirable for pipe-line use. It is strongly recommended that a system such as employed in this paper be followed as standard practice to facilitate the interpretation of diagrams.

Appendix. Device Designations (Based on ASA Standard C37.2)

1	Master element (start, stop, close, trip)
4	Master relay
20SC	Suction valve closing contactor
20SO	Suction valve opening contactor
20S1	Limit switch open only when suction valve is fully open
20S2	Limit switch closed only when suction valve is fully open
20S3	Limit switch open only when suction valve is fully closed
20DC	Discharge valve closing contactor
20DO	Discharge valve opening contactor
20D1	Limit switch open only when discharge valve is fully open
20D3	Limit switch open only when discharge valve is fully closed
20D4	Limit switch closed only when discharge valve is fully closed
20V	Vent valve
26P	Pump case overtemperature thermal device
30	Annunciator relay
	30-1, etc., individual drops
	30-O operate coil
	30-R reset coil
38M1	Motor bearing thermal relay (inboard)
38M2	Motor bearing thermal relay (outboard)
38P1	Pump bearing thermal relay (inboard)
38P2	Pump bearing thermal relay (outboard)
42	Motor circuit breaker
	a —Auxiliary switch closed when breaker is closed
	aa —Auxiliary switch closed when operating mechanism of breaker is in energized or operating position
	b —Auxiliary switch open when breaker is closed
	CC—Closing coil
	TC—Trip coil
	X —Closing relay
	Y —Closing cutoff relay
46	Phase current balance relay

47	Undervoltage or reverse phase voltage relay
	L—Contact closed for abnormal conditions
	N—Contact closed for normal conditions
47X	Auxiliary undervoltage for reverse phase voltage relay
48	Incomplete sequence timing relay
48X	Auxiliary incomplete sequence relay
49	Motor thermal overload relay
51S	Station overcurrent relay
51	Motor overcurrent relay
52	Main circuit breaker
	a —Auxiliary switch closed when breaker is closed
	aa —Auxiliary switch closed when operating mechanism of breaker is in energized or operated position
	b —Auxiliary switch open when breaker is closed
	CC—Closing coil
	TC—Trip coil
	X —Closing relay
	Y —Closing cutoff relay
52A	Auxiliary power breaker
52N	Auxiliary segregating breaker
	TC—Trip coil
63L1	High main sump level switch
63L3	Seal failure contact
63P1	Low suction pressure switch
63P2	High pump case pressure switch
63P3	Pump differential pressure switch

63P5	High discharge pressure switch
63P5X	High discharge pressure auxiliary relay
80	Battery undervoltage relay
86S	Station lockout relay
86	Pump unit lockout relay
88BC	Blower fan starter contactor
88EC	Exhauster fan starter contactor
88F	Motor air flow switch
88FX	Auxiliary timing relay for device 88F
96	Low instrument air switch
97	Low room air differential pressure
98	Hazardous atmosphere protective device
A	Ammeter
V	Voltmeter
W	Wattmeter

References

1. ELECTRICITY'S CONTRIBUTION TO MODERN PIPELINING, M. A. Hyde. *Westinghouse Engineer*, East Pittsburgh, Pa., Sept. 1953, pp. 154–61.

2. DESIGN OF AN UNATTENDED AUTOMATIC BOOSTER STATION, J. P. Hedlin. *World Oil*, Houston, Tex., Jan. 1951, pp. 188–90.

3. APPLICATION OF PIPELINE CONTROLS TO FIELD OPERATIONS, H. T. Chilton, Jr. *Oil and Gas Journal*, Tulsa, Okla., March 3, 1952, pp. 85–89.

4. CASE HISTORY AND ECONOMICS OF AN AUTOMATIC BOOSTER STATION, G. P. Jennings, M. E. Worlow. *Ibid.*, June 15, 1953, pp. 113–16.

5. REMOTE OPERATION OF PIPE-LINE PUMPING STATIONS, W. A. Derr, M. A. Hyde. *AIEE Trans-actions*, vol. 73, pt. II, Sept. 1954, pp. 190–98.

6. FOUR UNATTENDED PIPE-LINE PUMPING STATIONS, R. S. Cannon, T. V. Bockman. *Oil and Gas Journal*, Tulsa, Okla., Dec. 7, 1953, pp. 101–03.

7. NEW DEVELOPMENTS IN REMOTE CONTROL OF PUMPING STATIONS, M. A. Hyde. *Ibid.*, Oct. 13, 1952.

8. REMOTE OPERATION OF OLDER STATIONS, Dean Hale. *Petroleum Engineer*, Dallas, Tex., Sept. 1953, pp. D4–D6.

9. MICROWAVE COMMUNICATION, N. B. Tharp. *Westinghouse Engineer*, East Pittsburgh, Pa., Nov. 1952, pp. 198–203.

10. MICROWAVES FOR PIPELINERS, PART VIII, SUPERVISORY CONTROL AND TELEMETERING, C. B. Lester. *Pipe Line News*, Bayonne, N. J., July 1953, pp. A1–A8.

11. LEASED CHANNELS FOR METERING AND CONTROL, H. H. Joyner. *Electric Light and Power*, Chicago, Ill., Dec. 1953, pp. 100–06.

12. REMOTE MEASUREMENT AND CONTROL, W. E. Rufleth. *Petroleum Engineer*, Dallas, Tex., Oct. 1952, p. D-20.

13. AN ELECTRIC TELEMETERING SYSTEM, Carl Oman. *Radio and Television News*, Radio-Electronic Engineering Edition, Chicago, Ill., Oct. 1952, pp. 14–16.

14. A HIGH-SPEED TELEMETERING SYSTEM WITH AUTOMATIC CALIBRATION, W. E. Phillips. *AIEE Transactions*, vol. 70, pt. II, June 1951, pp. 1256–60.

15. NEW ELECTRICAL PRESSURE INSTRUMENTATION ON PETROLEUM PIPE LINES, M. J. Dabney. *Proceedings*, National Conference on Industrial Hydraulics, Chicago, Ill., vol. 5, 1951, pp. 195–202.

16. AUTOMATIC STATION CONTROL, SUPERVISORY, AND TELEMETERING EQUIPMENTS. *ASA C37.2-1945*, American Standards Association, New York, N. Y., Dec. 11, 1945.

Eddy-Current Press Drives

F. L. HOPF
ASSOCIATE MEMBER AIEE

T. R. LA VALLEE
NONMEMBER AIEE

THE perennial requirement of more pieces per hour from all presses in a modern metal stamping plant has created many problems. Obviously, every press must operate at a greater number of strokes per minute with correspondingly higher punch velocity. This situation is aggravated by variations in steel sheets both from a chemical analysis and thickness consideration.

These several problems instigated the development of better methods of driving and controlling presses. One system uses a direct-connected motor which starts and stops on each stroke. This drive is limited to the torque that the motor can develop, since a flywheel is not used. As a result special high-torque motors are employed.

This situation resulted in the application of an eddy-current clutch and brake as a press drive. The use of this combination which incorporates a flywheel permits selection and control of punch velocity anywhere in the stroke with a normal size standard induction motor. Models are available for single-, double-, and triple-action presses rated 300 to 1,500 tons. The basic design, application, and excellent performance of these units is described in this paper.

Speed-Torque Characteristics

In the application of an eddy-current drive for mechanical presses, two basic problems are of prime importance. The first of these is the generation of sufficient torque for the job to be done. The second is to ensure adequate dissipation of heat which occurs in the functioning of the press. Fig. 1 illustrates the basic speed-torque relationship used for both the clutches and brakes of a typical press drive. Both units develop their rated torque at a specified slip speed. As noted, both also develop about 10-percent (%) excess torque above the rated value. This serves two purposes: the excess torque is advantageous for accelerating the driven system of a press when operated intermittently and by limiting the excess torque to about 10% some safety against breakage is engineered into the application. The uppermost curve in Fig. 1 illustrates the capacity of both the clutch and brake at rated excitation. At any reduced value of field excitation, the same characteristic curve is available but at a lower level of torque for the corresponding slip values. In this respect, the curve is quite similar to an induction motor designed for high slip. The utilization of an eddy-current press drive does not, however, permit the use of a standard slip induction motor to drive the press, particularly on large, slow-speed, double- or triple-action machines. It does, however, permit on large, slow, draw presses the flywheel to give up its energy during the draw while transmitting substantially constant torque.

While Fig. 1 shows a typical torque-speed curve of an eddy-current press drive clutch or brake at rated excitation, obviously lesser amounts of excitation will show a characteristic reduced in ampli-

SLIP SPEED RANGE DURING COMPLETE PRESS CYCLE
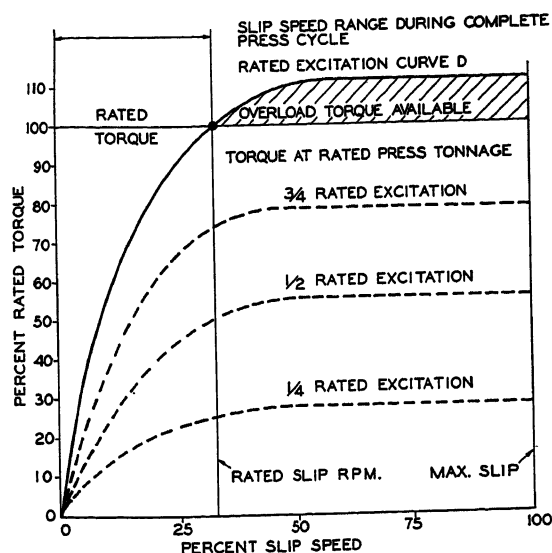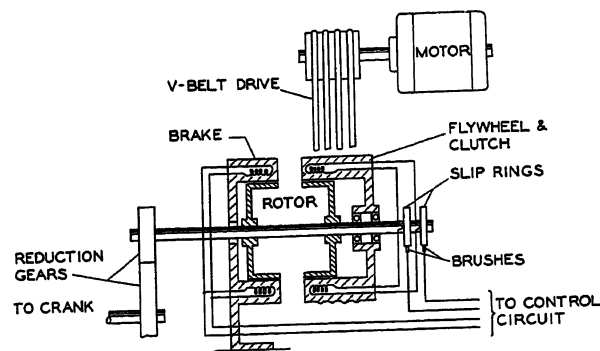
RATED EXCITATION CURVE D

Fig. 1 (left). Typical torque-speed curves of clutch or brake in press drive

Fig. 3 (right). Schematic eddy-current clutch and brake

tude but of similar shape. Thus, the clutch may be operated continuously at a fixed value of excitation or particularly on slow-speed presses; the amount of excitation can be varied at different points during the stroke to achieve speed variation when required. This is of considerable value as will be seen later.

Heat Characteristics and Cooling

As has already been pointed out, eddy-current machinery must, at any excitation point, slip to develop torque; therefore, the speed differences between the input and output multiplied by the torque transmitted represent power loss and are rejected in the form of heat. The magnitude of the slip loss for any speed and load point is calculated by the following equation

$$\text{horsepower (loss)} = \frac{\text{slip rpm}}{\text{output rpm}}$$
$$\text{horsepower (load)}$$

With a constant torque load the resultant

characteristic is a straight line in which the losses are directly proportional to slip. However, if the torque load is reduced, the operating losses are decreased directly in proportion to the reduction in torque.

Because of the highly cyclic load known to exist in a mechanical stamping press, the losses are of considerably smaller order than would be thought at first glance. The worst condition is obviously when starting up the press from a cold condition or for intermittent operation where it is necessary to accelerate the entire driven system initially or for each stroke. Obviously, the slip is 100% at the time excitation is applied to the field coil. Experience, however, indicates that most of these drives will accelerate the driven system of the press in about 2 revolutions of the clutch rotor and from a time standpoint from 0.5 to 1.0 seconds, depending upon the size of the press. Heating also occurs when the speed of the output or driven system of the press is arbitrarily arrested to some degree by

the eddy-current brake while the clutch portion is de-energized, then reaccelerated by reintroducing field excitation to the clutch coil. The work portion of the cycle rarely exceeds one-sixth of the total stroke. Consequently, if the clutch slips the maximum of 30% during this period when averaging over 360 degrees, it is a very small amount. A study of a large number of presses so equipped, rated between 300 and 1,500 tons, indicated that the heat dissipation of the clutch is not a serious problem. In fact, in some cases the slip loss during the draw was negligible.

From the foregoing the feasibility of air-cooling these units is evident. By incorporating a built-in fan, cooling air is directed over the finned clutch and brake rotors and exhausted at the flywheel. The merit of this arrangement has been demonstrated by successful operation of several hundred units.

Construction of an Eddy-Current Press Drive

The first eddy-current press drives were mounted on the side of the flywheel in the same manner as other clutches in use at the time. While a number of these are in daily operation, it was felt that a
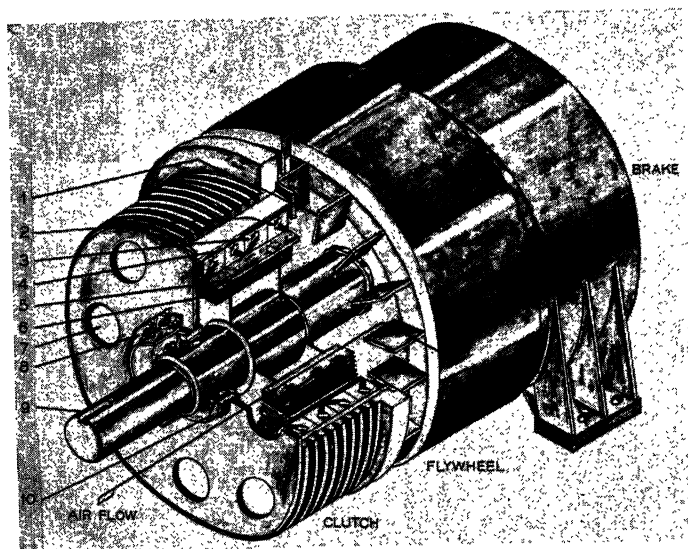

Fig. 2 (left). A large eddy-current press drive

1. Center discharge fan
2. Form wound coil
3. Individual field assembly
4. Clutch housing
5. Clutch rotor rim
6. Rotor fin
7. Ventilation holes
8. Slip ring assembly
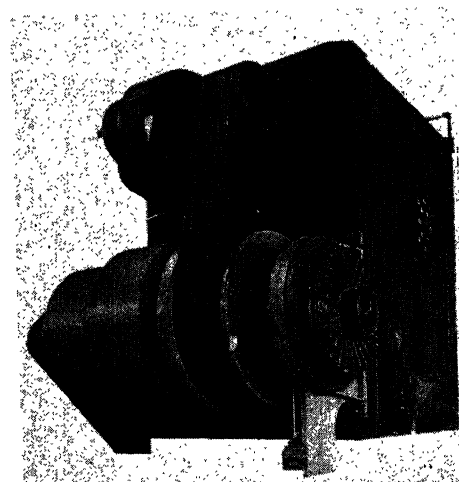9. Output shaft
10. Heavy - duty roller bearing

Fig. 4. Press drive application on a 1,500-ton press
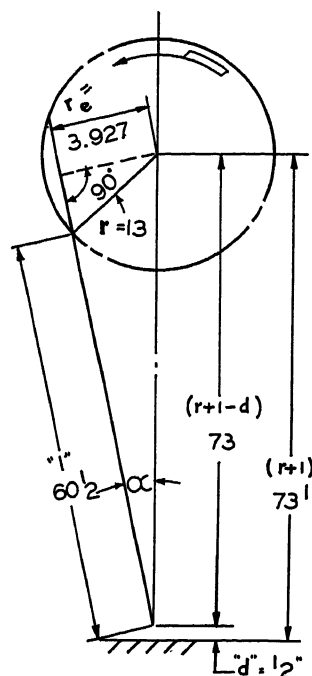
$$R = \frac{\text{Flywheel RPM}}{\text{Strokes/Min.}} = \frac{419}{12}$$
$$= 34.95$$

r_e = Effective Arm

$$= \text{Sine } \alpha C \times (r + 1 - d)$$

$$= .05379 \times (73)$$

$$= 3.927$$

$$\text{Cosine } \alpha C = \frac{(r+1 - d)^2 + 1^2 - r^2}{2 \times 1 \times (r+1 - d)} = \frac{73^2 + 60\frac{1}{2}^2 - 13^2}{2 \times 60\frac{1}{2} \times 73} = \frac{8820.25}{8833} = .99878$$

Angle $\alpha C = 3° - 4$ feet - 44 inches (Handbook)

Sine $\alpha C = .05379$ (Handbook)

$$\text{Torque Reqd. (lbs. ft.)} = \frac{\text{Tonnage (at 'd' inches)} \times 2000 \text{ lbs./ton} \times r_e \text{ (inches)}}{12 \text{ inches/Ft.} \times R \text{ (torque multiplier ratio)}}$$

$$= \frac{500 \times 2000 \times 3.927}{12 \times 34.95} = 9370 \text{ lbs. ft.}$$

Fig. 5. Press drive calculation sheet. Drive on top of press

redesign to place the clutch portion of the drive within the flywheel offered a number of advantages. In Fig. 2 a cross section of a modern unit rated at 90,000 pound-feet torque at 400-rpm clutch capacity and 60,000 pound-feet at 400-rpm brake capacity is shown. This clutch is composed of six separate magnetic sections. Each of the field sections consists of a toroidal coil surrounded by cast pole sections and securely bolted together. The six are keyed to the rim-type flywheel and operate, for all practical purposes, independently of each other. These field coil assemblies are identical in the clutch and in the 4-coilbrake unit. At either side of the ring-type flywheel side plates retain the coil sections in place and provide for the bearing support of the rotor. The brake housing is very similar except that it is arranged to anchor directly to the bed of a press. The rotors of both the clutch and brake have very thin rims, heavily finned for heat dissipation, and are supported by thin side members attached to a central hub. Thus, the iner-

tia of the rotors is kept to an absolute minimum which is essential in rapid response of the driven system of the press which is ordinarily a requirement. All clutch brake coils are connected in parallel and the clutch coil leads are brought out through thyrite resistors for inverse voltage protection to four slip rings mounted on one of the side plates. Each of these rings is fed by two brushes located at 180 degrees to minimize electric disturbances from vibration or foreign material on the surface of the rings. The coil leads of the brake are brought out to a junction box for connection to the press electric system.

Stored energy was a particular problem in this design. This accounts for the extremely heavy ring-type flywheel and the addition of similar rings inside the flywheel adjacent to clutch coil sections. This particular unit has a stored energy at 400 rpm of nearly 3,000,000 foot-pounds. The design operating point for the torque rating was developed at a slip speed of 90 rpm or 22.5%. A propor-

tionate amount of torque was available in the brake at the same slip speed.

In Fig. 3 the elemental press drive shown illustrates how such a unit is attached to a press showing the rest of the connections in a line diagram. This best illustrates the general application. Fig. 4 illustrates the 1,500-ton press utilizing the drive shown in cross section in Fig. 2.

Selection of Clutch Capacity

While there are several methods of determining the proper clutch size for a given press requirement, the most satisfactory for sizing eddy-current equipment is shown in Fig. 5. The information shown at the top of the figure adjacent to the force diagram is normally supplied by the press manufacturer to meet a definite requirement. As shown in the lower half of the calculation, the given values and application of the cosine law can easily be used to determine the torque capacity required of the clutch at a given distance from the "die-closed" position. The calculation in Fig. 5 is for a 500-ton eccentric toggle double-action unit. The next larger standard size press drive was used, namely a 10,500-pound-foot unit. This proved entirely satisfactory even when used much more severely than was originally planned.

Selection of Brake Capacity

Normally, an eddy-current air-cooled brake is used to complement the clutch for a complete drive for a mechanical press. The use of the brake, however, does not obviate the requirement for a full-capacity, spring-loaded, friction brake, usually actuated by air. Practically all states require by law such a brake in the interest of safety. The eddy-current air-cooled brake furnished with the PD-10,500 clutch just described utilized two of the same coil sections and rotors in the same diameter. The primary reason for selecting a brake of lower capacity than the clutch is that on this application it has the advantage of operating at 100% slip at the time it is energized. Thus, greater torque capacity can be obtained per coil section than from the clutch which normally operates at 20 to 30% slip. Most of the retarding effort is accomplished by the eddy-current brake, thereby prolonging the life of the friction brake. Experience to date indicates that brakes for such units can normally be used with a rating of 25 to 50% of that of the eddy-current clutch. On heavy-duty, slow-speed presses operated intermittently, the brake duty is most severe inasmuch
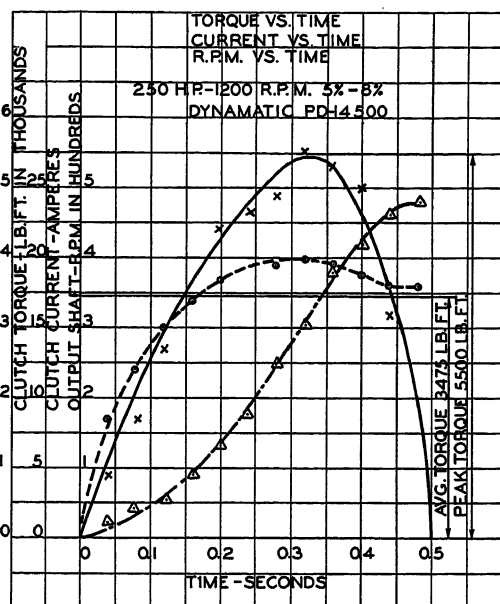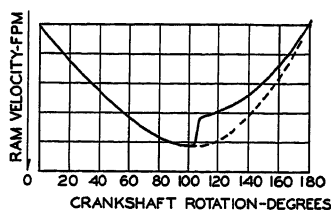
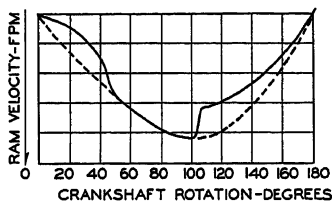Fig. 6. Acceleration test data on large press

Fig. 7. Normal velocity ram operation, full excitation

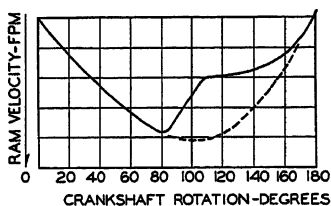Fig. 8. Normal velocity ram, intermittent operation, full excitation

Fig. 9. Half velocity ram, full excitation

as the press must be stopped at each stroke. Using this basic information for design a large number of presses rated up to 1,000 tons and others rated up to 14 strokes a minute are in daily use.

Press Operation with an Eddy-Current Drive

As is the case in many other problems, there is a definite period of time between when excitation is applied to the slip rings of an eddy-current clutch used as a press drive and the time that field excitation is built up to a point where useable torque is obtained. While the direct current supplied to the coil for rated excitation is a small percentage of the transmitted horsepower, the time interval is worthy of some consideration. The usual manner of expressing this situation is with a time constant defined as the period of time required for obtaining 63% of the rated current through the coil. As previously mentioned, a typical drive of this type for a 700- or 800-ton press will accelerate within about 2 revolutions of the rotor and in about 0.5 second. Such a curve is illustrated in Fig. 6. These curves show the current build-up in the coil, the rpm build-up in the output shaft, and the torque build-up in that shaft. It is noteworthy that the time interval for acceleration is so small. This curve is an actual test without field forcing. Further tests proved that field forcing would reduce the time interval. Other test data on a 1,000-ton machine were of the same general type but the time for acceleration without field forcing was increased to about 0.9 second.

It has already been pointed out that a press does its work in 1/6th revolution or less, which represents the working portion of the stroke on a double- or triple-action draw operation. Therefore, the only concern is with the motion of the press slide between 90 and 180 degrees, from top dead center, usually between 120 degrees, and when the dies close. As mentioned in the foregoing, the eddy-current clutch of a press drive must develop slip before developing torque. Therefore, after the press accelerates to normal speed it maintains that speed until the punch contacts the metal for a draw operation or any other. At this point,

the clutch rotor changes speed very rapidly until sufficient torque is developed to overcome the work in the press. This is shown in Fig. 7; the curve illustrates operation with an eddy-current clutch at full excitation with the press operated "on the hop." The point at which the slip occurs depends upon the type of operation being done and may occur to 180 degrees. If the same press and drive are operated on an intermittent basis, the velocity line appears as shown in Fig. 8 which includes the acceleration period required at the beginning of each intermittent stroke. In each case, the dotted line represents a "geared line" such as would result if the input and output shafts were rigidly connected. The magnitude and duration of the slip during acceleration and during the work portion of the cycle are entirely dependent upon the press and the operation being performed. This slip characteristic cushions the punch as it goes into the work and allows the metal a little bit more time to flow as it is drawn, thus minimizing scrap material resulting from a fast punch entry.

Of considerably more value is the type of operation shown in Fig. 9. This type of operation results from interrupting the clutch excitation at a given point of crank shaft motion and simultaneously supplying brake excitation. The speed of the output shaft of the clutch is reduced very rapidly, thereby reducing the velocity of the punch to any predetermined value to about 30% minimum. At the time that the punch reaches the desired value of velocity, the brake excitation is removed and the clutch re-energized, thus completing the draw. The high speeds at which presses are being geared today make this feature of paramount importance. The characteristic curve exhibited in Fig. 9, to be better understood, is redrawn as shown in Fig. 10, which illustrates the complete press cycle. Here the slow
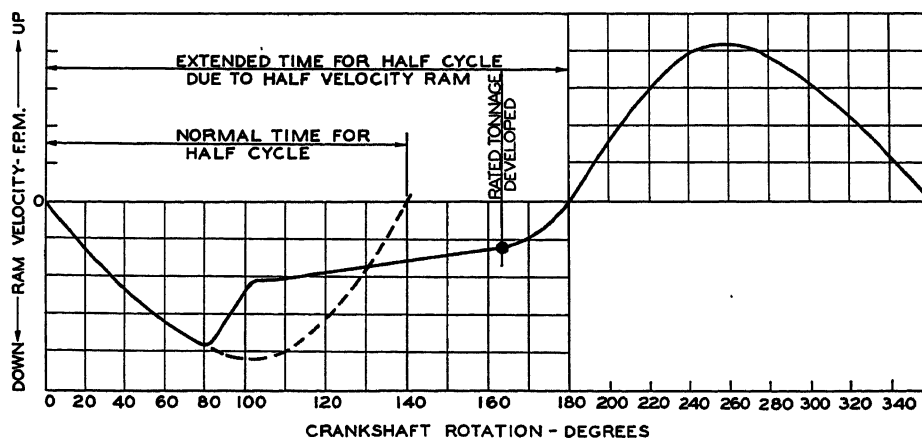

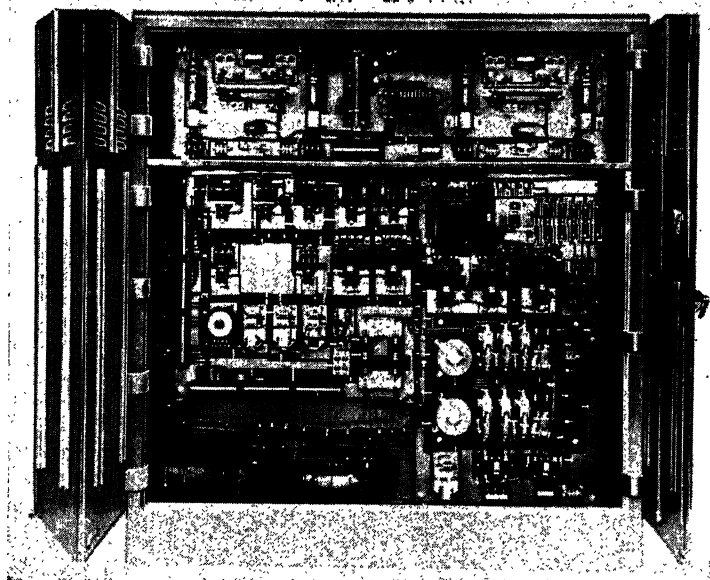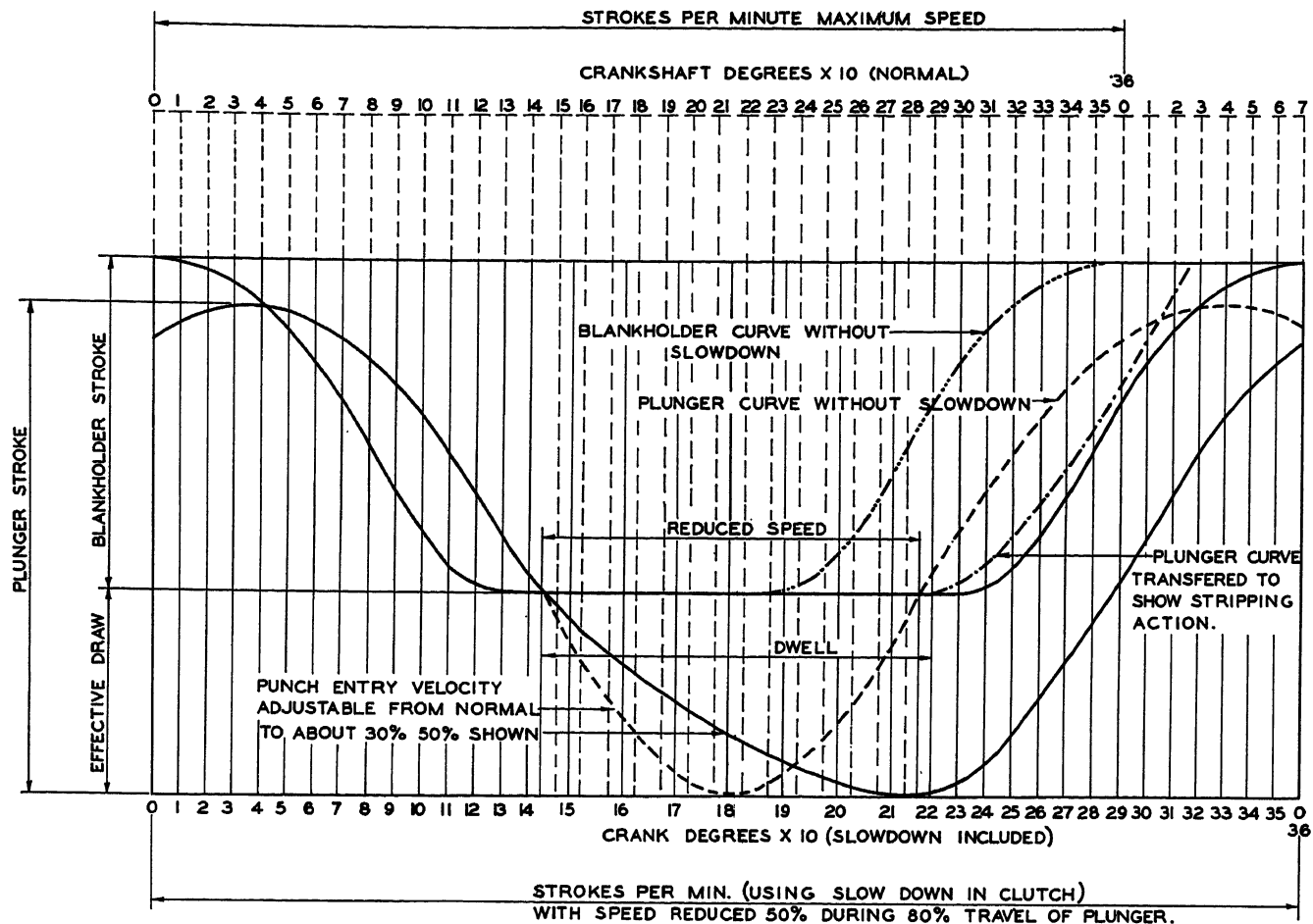Fig. 10. Compositive curve, half velocity, ram for drawing cycle

STROKES PER MINUTE MAXIMUM SPEED

CRANKSHAFT DEGREES X 10 (NORMAL)

BLANKHOLDER CURVE WITHOUT SLOWDOWN

PLUNGER CURVE WITHOUT SLOWDOWN

REDUCED SPEED

PLUNGER CURVE TRANSFERED TO SHOW STRIPPING ACTION.

DWELL

PUNCH ENTRY VELOCITY ADJUSTABLE FROM NORMAL TO ABOUT 30% 50% SHOWN

CRANK DEGREES X 10 (SLOWDOWN INCLUDED)

STROKES PER MIN. (USING SLOW DOWN IN CLUTCH)
WITH SPEED REDUCED 50% DURING 80% TRAVEL OF PLUNGER.

Fig. 11 (above). A complete double-action press cycle with slow down

Fig. 12 (left). A typical press control panel

down of 50% has been plotted to a time scale thereby indicating the almost constant draw throughout the critical portion of the work stroke. This curve indicates a press operation "on the hop." This same motion is transferred to a double-action press motion curve, Fig. 11. If the press is operated intermittently, an acceleration portion would also show in

Fig. 10 as in Fig. 8 between zero and 40 degrees.

Utilizing this principle, presses have been operated with the punch stopped completely at a point where normally it would have maximum velocity; then the eddy-current clutch was re-excited and satisfactory stampings were drawn. This severe test was used for demonstration

purposes only, but it illustrates the flexibility of this operation. This principle in varying degrees is in everyday use by large automotive manufacturers.

Control for Eddy-Current Press Drive

It is evident from the foregoing that the precise operation of an eddy-current press drive, particularly as shown in Figs. 10 and 11, is contingent on accurate timing and precise voltage control of the power supplied to the clutch and brake. Fig. 12 illustrates a typical excitation supply panel used in conjunction with an eddy-current press drive. Basically it consists of separate thyratron rectifiers for the clutch and brake; the clutch rectifier grid-phasing network; the brake rectifier grid-phasing network; and a relay panel and a cam-operated limit switch (not shown). The phasing circuits of both the clutch and the brake rectifiers are so arranged that when all resistance is cut in the electronic tubes invert and rapidly discharge the coil. To provide for varying amounts of excitation for the clutch winding, the resistor of the clutch phasing circuit is tapped. The relay panel, in conjunction with the rotary cam limit switch (not shown), sets up and ad-

justs the phasing circuits of the clutch and brake and provides safety features. Some of the later control panels for this purpose incorporate current limiters, thus affording forcing voltage to minimize the time constant of the eddy-current clutch and brake for more rapid response.

Conclusions

In addition to the fundamental advantages of no friction surfaces, in the usual sense, vibration isolation, and low maintenance, these units now provide a new tool for use in conjunction with modern high-speed mechanical stamping presses. While it is possible through the use of direct-connected motors to achieve slow draw and speed variations within the drawing portion of the stroke, such op-

eration is limited to motor torque and is accomplished at relatively high cost. A similar operation can be obtained from a hydraulic press, but the initial cost is higher and the production rates are somewhat less.

Eddy-current press drives are available in sizes from 7,000 pound-feet at 550 rpm to 120,000 pound-feet at 425 rpm. Similar brakes utilizing the same coils and field sections are available to match the clutch capacities. Smaller capacity units have been offered in the past and probably will be offered again in the future. When they reappear, they will utilize the advantages pointed out in the latest construction methods described in this paper. It is possible through the medium of toggle motion to arrange a mechanical press for slow down and fast recovery. Any such

fixed system has the inherent disadvantage of requiring almost a complete rebuilding job to alter the degree and rate or slow down. With an eddy-current press drive, the draw speed can be changed in a very short period of time to match the metal being processed or the job being placed in the press. For job shop operation the latter feature is particularly important as conditions vary over a wide range.

From a power consumption standpoint the use of the eddy-current press drive requires no increase in motor size. The negligible slip occurring from capacity use of the press does not reduce the production rates. Continued research and development will add many new uses for mechanical presses to the types of operation outlined.

Electrical Grounding and Cathodic Protection at the Fairless Works

W. E. COLEMAN
ASSOCIATE MEMBER AIEE

H. G. FROSTICK
MEMBER AIEE

A STEEL grounding system has been installed at Fairless Works, the new plant of the U. S. Steel Corporation near Morrisville, Pa. The use of steel instead of copper for electrical grounding was adopted as part of a program to minimize underground corrosion. Since this is a departure from conventional practice, it is believed that the story of this particular installation will be of interest to many engineers.

During the planning stage of this new plant, research engineers recommended that the underground structures be designed to provide as much protection against underground corrosion as economically possible. The recommendations included the following main suggestions:

1. All bare copper underground should be eliminated.

2. All buried pipe lines should be coated.

3. An insulating covering should be used on all lead cables.

4. Steel rods containing 6-per-cent chromium should be used for electrical grounding.

5. Cathodic protection should be used to prevent corrosion of the steel ground rods and to provide some measure of protection for other necessarily exposed steel or lead.

Because of procurement difficulties, the

6-per-cent chrome steel was finally replaced with plain carbon steel. This was not believed to be detrimental, because cathodic protection was included in the plan.

After these recommendations had been accepted in substance by the design engineers, it became necessary to lay out the grounding system on a practical basis. Accordingly maximum ground-resistance values were established:

1. At 69-kv substations, 1 ohm

2. At transmission-line towers, 11 ohms

3. At 13.8-kv substations and at building ground beds, 3.5 ohms

4. At 2.3-kv substations, 7.5 ohms

5. At isolated buildings, roadway-lighting standards, etc., 25 ohms

The electric system at Fairless Works is designed for grounding all the a-c power sources. There are four a-c potential levels: 69 kv for open-wire transmission lines; 13.8 kv for underground distribution and for operation of motors about 2,000 horsepower and larger; 2.3 kv for utilization in power blocks between 100 and about 2,000 horsepower; and 440 volts for equipment of 100 horsepower and smaller. Lighting circuits are energized at 110 volts from a 440-volt pri-

mary supply. In addition, there is a 250-volt d-c power supply for crane and constant-potential auxiliary services.

All power sources are standardized. The transformers furnishing 13.8-kv power from the 69-kv primary are sized at 33,000 kva; 2,300-volt power is furnished through transformers rated 2,500 kva; and 440-volt power is supplied by 1,500-kva units. The constant-potential direct current is furnished almost entirely by 1,500-kw mercury-arc rectifiers.

All transformers have delta-connected primaries and wye-connected secondaries. The 69-kv system is solidly grounded. The 13.8-kv systems are grounded at the transformer neutral through 10.7-ohm resistors. The 2,300- and 440-volt transformers are solidly grounded at the wye-connected secondaries.

There are exceptions to these transformer sizes as well as exceptions to the practice of grounding the transformer neutral. One example of the latter is the auxiliary power supply in the powerhouse, for which the two sets of transformers have sufficient capacity so that either set may carry the entire load. Ground indication only is furnished at this point. Another example is at the pump house, where two 3,750-kva transformers (69 kv to 2.3 kv) are connected to ground through 2-ohm resistors that (by means of a man-

ually operated disconnect switch) may be removed from the circuit. At this second location, if a ground-fault relay disconnects a piece of equipment from service, and if the fault current has not damaged the effectiveness of the equipment, the control operator may take the calculated risk of removing the ground resistance and continuing to operate the apparatus. This may be done only when service must be maintained regardless of possible damage to apparatus.

Grounding these various power supplies offers many advantages, the most important of which is the improved protection to both personnel and equipment. This protection is achieved by fixing the potential to ground, by limiting the ground-fault current, and by utilizing high-speed, positive, ground relaying. Since greater gain is to be achieved by use of a ground connection, careful attention must be given to the means by which the electric system, buildings, and other underground structures are connected to earth. As may be inferred from this brief description, the electric system at Fairless Works is quite extensive.

Grounding Requirements

The over-all electrical plan called for 5 outdoor substations, 14 indoor substations, 1 power house, 45 building ground beds, 6 miles of overhead power lines, and 20 miles of underground multiple duct-line. Fig. 1 shows the location of ground beds in the portion of the plant devoted to sheet and tin production. It was first suggested that the electrical grounding be so placed at each location that it could be completely isolated geometrically from the installation it was protecting. This proposal had to be modified in the substation locations, where it was necessary to distribute the grounding system to provide adequate safety for personnel. Grounding was not planned for individual manholes, but all cable sheaths were to be bonded together at every third manhole

and the sheaths grounded at terminal points.

The size and number of ground rods and the type of cathodic protection used were based on a field survey of the site to determine soil resistivity and composition. The first field measurements were made in May 1951, and were confined to measuring soil resistivity. The method of measurement was the 4-electrode system, in which a 4-terminal ground-resistance megger was used. The values obtained ranged from a minimum of 3,600 ohm-centimeters to a maximum of 90,000 ohm-centimeters. The average value was approximately 30,000 ohm-centimeters, a value considerably higher than had been expected. Calculations based on this figure indicated that to obtain a 1-ohm ground would require 100 steel rods (1 inch in diameter), driven 40 feet into the earth. The high resistivity of the soil also indicated that rectifiers would probably be desirable for cathodic protection. However, it was reasoned that even though the ground resistance was high, it would be much more desirable to use a form of cathodic protection that would not require special power connections and would not require maintenance at frequent intervals. Consultation with suppliers of magnesium resulted in a decision to use specially designed magnesium anodes for the cathodic protection. The anode recommended was made up of a 1-inch-diameter steel rod 20 feet long, with a piece of magnesium 10 feet long and 4 inches square cast on the center section of the rod; see Fig. 2. This arrangement permits using the anode as an additional grounding electrode, as well as an electrode for cathodic protection. The recommended plan for installing the anodes was to place them in a hole 15 feet deep and 8 inches in diameter or 10 inches square, and then to backfill with a wet mixture consisting of 50-per-cent bentonite, 25-per-cent gypsum, and 25-percent sodium sulfate.
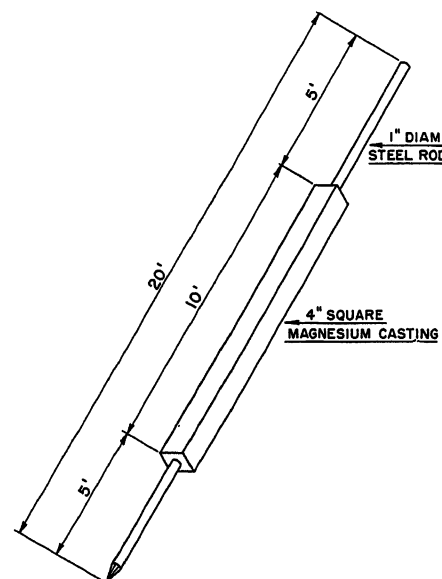
The basic plan called for the placement

Fig. 2. Sketch of magnesium anode

of the steel rods and the magnesium anodes in the arrangement shown in Fig. 3. The design called for the use of one insulated copper cable to connect all the steel rods and a second insulated copper cable to connect all the magnesium anodes. Each of these cables was then to be connected separately to the building frame or ground bus above ground level. In this manner, the magnesium or the steel could be disconnected easily for measurement purposes. Variations of this basic ground-bed arrangement were made, as necessary, to provide proper resistance values and to suit space requirements. In the substation locations, it was necessary to modify the basic scheme considerably and to distribute the ground rods and anodes in a manner that would provide for maximum safety. Fig. 4 shows the arrangement used at the main
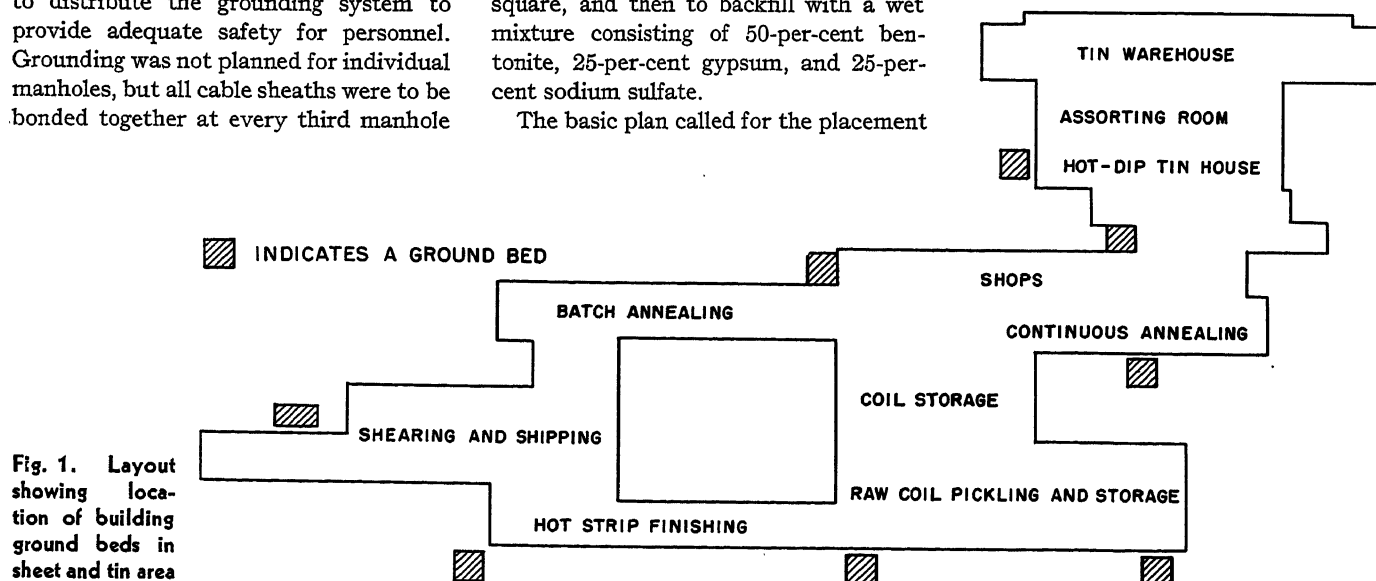

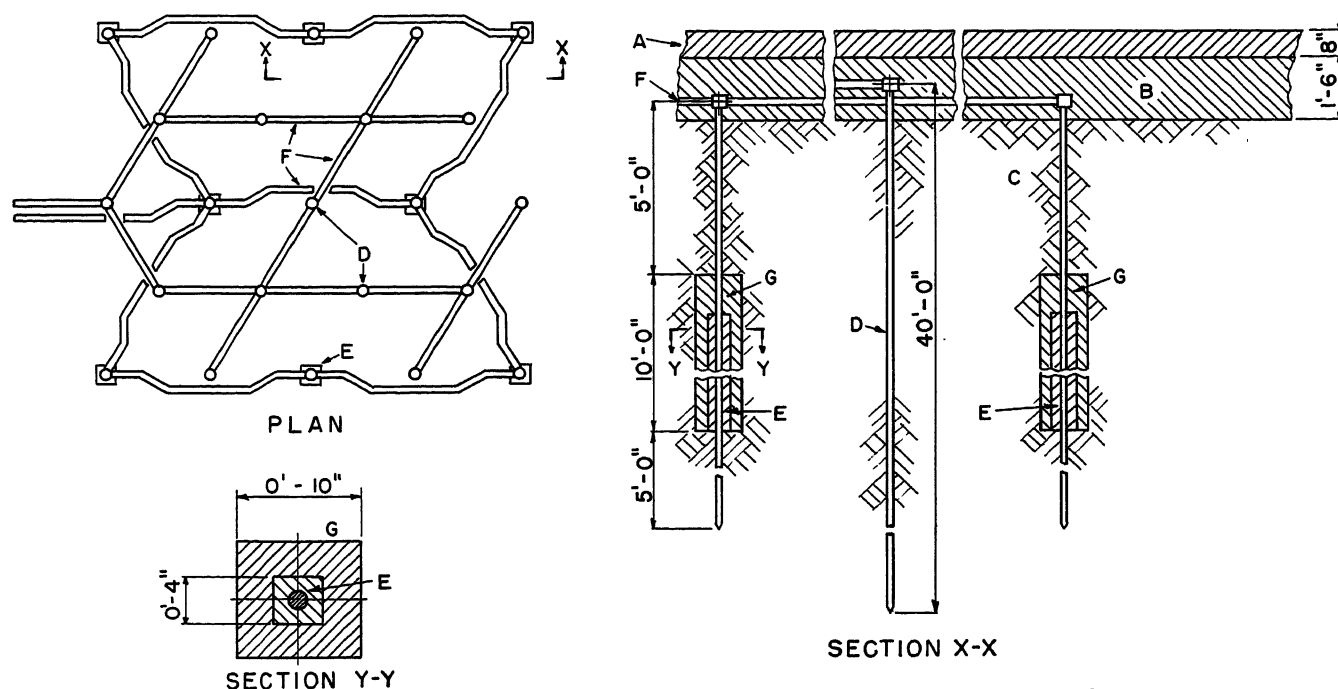**Fig. 1.** Layout showing location of building ground beds in sheet and tin area

PLAN

SECTION Y-Y

SECTION X-X

Fig. 3. Basic ground-bed arrangement

A—3/8-inch crush rock
B—Crushed rock
C—Soil
D—Ground rod (low carbon steel)
E—Magnesium anode
F—Copper cable (insulated)
G—Backfill: 50-per-cent bentonite, 25-per-cent gypsum, and 25-per-cent sodium sulphate

substation, this arrangement being representative of those used at all substations.

During the time the final design was being developed, additional field tests were conducted to confirm the calculations as well as to verify the design. A test site was chosen about 200 feet east of the proposed location for the main substation. This point was on virgin soil, at an elevation of 18 feet. Soil resistivity was measured and found to average about 40,000 ohm-centimeters. The actual values measured are given in Table I. The M. C. Miller corrosion-test set and the megger were both used, and the results checked within 5 per cent.

To determine ground resistance and to measure cathodic-protection currents, a magnesium electrode was buried in a prepared backfill, and several steel rods were driven to various depths. The electrode consisted of five 21-inch anodes placed end to end and buried in a hole 15 feet deep and 7 inches in diameter beside a steel rod. A wire from each anode led to the surface so that individual currents might be measured. This arrangement, used in lieu of the proposed anode shown in Fig. 2, which was not yet available, permitted the determination of cathodic-

current flow at various depths. The steel rods were 1 inch in diameter and were driven in 10-foot lengths, the lengths being welded together as needed. A compressed-air jack hammer was used to drive the rods. After some difficulty, one rod was successfully driven to 38.5 feet, and its resistance to ground was measured as 13.4 ohms, which checked very well with calculations. Because of

Fig. 4. Ground plan of main substation

LEGEND
● STEEL GROUND ROD
◉ MAGNESIUM ANODE
○ FENCE POST
△ CONNECTION POINT

25 FEET

the difficulty in driving the 1-inch rods, it was decided to change the specifications to provide 1¼-inch-diameter rods. This would not appreciably lower the ground resistance, but it would make the rods much stronger. They would have less tendency to bend under the hammer, and considerably more steel would be in the electrode structure.

After the electrodes had been allowed to

Table I. Soil Resistivity Readings at Test Site

Depth, Feet	Average Soil Resistivity, Ohm-Centimeters	
	1951	1953
10	23,900	23,000
20	38,200	32,500
30	44,200	31,500
40	44,200	28,300
50	47,600	
60	43,600	

Table II. Anode Currents at Test Site

	Average Current Output, Milliamperes		
	June 1951	May 1952	May 1953
Anode no. 1 (bottom)	24		6.4
Anode no. 2	20		5.5
Anode no. 3	22		11.8
Anode no. 4	24		18.0
Anode no. 5 (top)	20		12.2
All five anodes	32	36	27.0

Table III. Ground-Resistance Measurements at Four Ground Beds

	Resistance to Ground, Ohms					
	1952			1953		
	Magnesium	Steel	Magnesium +Steel	Magnesium	Steel	Magnesium +Steel
Building ground A	4.2	3.3	3.0			
Building ground B	5.8	3.3	3.1			
Building ground C	2.8	2.0	2.0			
Building ground D	6.25	3.0	2.8	3.5	2.2	2.0

Table IV. Soil-Potential Measurements at Typical Ground Bed (See Fig. 5)

Reference	Potential to Copper Sulphate, Volts			
	At Point A	At Point B	At Point C	At Point D
Building frame	−0.53	−0.44	−0.47	−0.465
Building frame plus steel ground rods	−0.60	−0.58	−0.51	−0.56
Steel ground rods	−0.77	−0.80	−0.78	−0.84
Building frame plus magnesium anodes	−0.82	−0.78	−0.56	−0.70
Building frame plus magnesium anodes plus steel ground rods	−0.85	−0.80	−0.57	−0.68
Magnesium anodes plus steel ground rods	−1.02	−0.88	−0.96	−1.10
Magnesium anodes	−1.35	−1.40	−1.30	−1.45

settle for 1 day, measurements were made on the cathodic-protection currents. The current measured from the anode to the 38.5-foot rod was 12.6 milliamperes, which is slightly more than 1 milliampere per square foot of steel. This value is generally accepted as sufficient for high resistivity soils. About 6 weeks later, on August 2, 1951, a further check was made at the same location, and no significant change was noted either in the resistance to ground of the 38.5-foot rod or in the protection current being furnished by the magnesium. The fact that this much current was being obtained in the high-resistance soil was at first a little surprising; therefore, soil samples were taken and laboratory tests were conducted to determine whether or not there could possibly be any error. The galvanic currents obtained in the laboratory with couples of

magnesium and iron in slightly damp soil checked with the field results reasonably well. An additional field test made in May 1952, nearly a year later, indicated substantially no change in the protection current, although the polarization potentials measured against a copper sulfate cell had changed somewhat (magnesium from 1.65 to 1.55 volts, and steel from 0.77 to 0.65 volt). In May 1953, a further check at this location showed some reduction in the total protection current. The current distribution among the anodes is given in Table II.

The results of the 1951 tests were considered sufficient support for the proposed design; therefore the installation was made in the spring of 1952 in accordance with the modified plans. The actual con-

struction turned out to be rather difficult in some of the locations because of the large pieces of gravel profusely distributed throughout the soil. In fact, one contractor experienced so much trouble in drilling holes for the anodes that the drilling rigs were abandoned and the holes were dug by hand. This resulted in holes with much larger diameters than were required, so that it was necessary to use more of the backfill than had been planned. In the entire installation, there are approximately 1,000 ground rods and 500 magnesium anodes.

After a portion of the installation had been completed, and before any of the plant was in operation, a series of tests was made in May and June of 1952 as a further check on the effectiveness of the
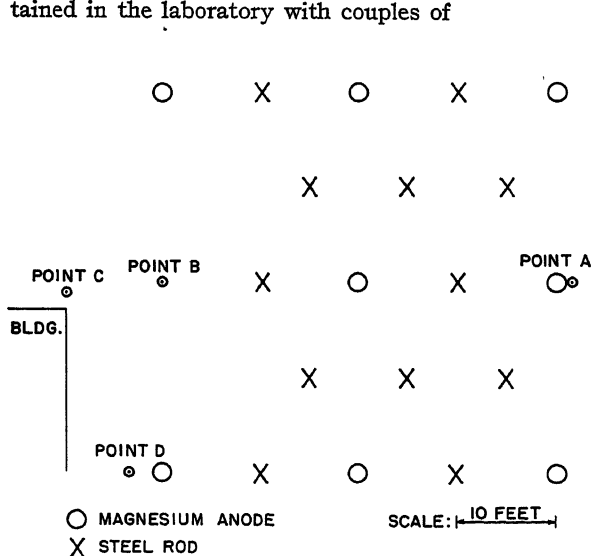

Fig. 5 (left). Ground bed at northwest corner of raw coil storage building showing location of soil-potential tests
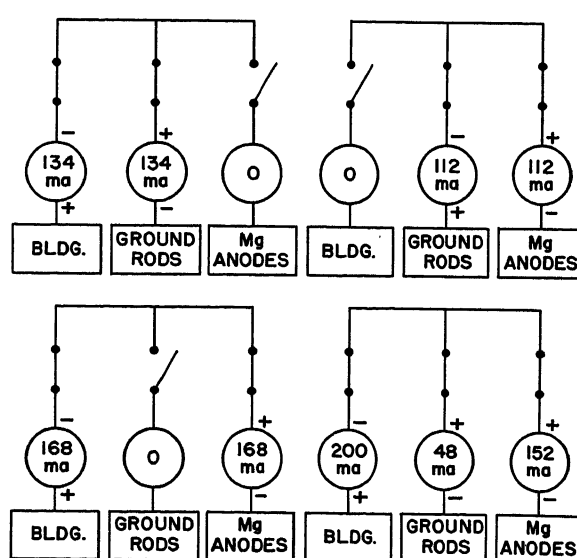
Fig. 6 (right). Galvanic current distribution — building ground bed at northwest corner of raw coil storage building

O MAGNESIUM ANODE
X STEEL ROD
SCALE: 10 FEET

Table V. Typical Data Sheet Showing Readings Taken at Building Ground Bed

| Location N.W. Corner, Raw Coil Storage Bldg. | Date 5/27/53 |
| Weather Fair—Cool—Wet | Made by W.E.C. |

1. Current between magnesium and balance of ground network at test plate*
   230 milliamperes magnesium polarity negative
2. Current between ground rods and balance of ground network at test plate*
   (a) With magnesium disconnected — 126–132 milliamperes ground rod polarity negative
   (b) With magnesium connected — 18 milliamperes ground rod polarity negative
3. Potential between ground rods and balance of network at test plate
   (a) With magnesium connected — 0.05 volt ground rod polarity negative
   (b) With magnesium disconnected — 0.32 volt ground rod polarity negative
4. Potential between magnesium and balance of network at test plate
   (a) With ground rods connected — 0.72 volt magnesium polarity negative
   (b) With ground rods disconnected — 0.90 volt magnesium polarity negative
5. Resistance to ground of magnesium — 3.0–3.5 ohms
6. Resistance to ground of ground rods — 1.8–2.2 ohms
7. Resistance to ground of magnesium and ground rods — 1.6–2.0 ohms
8. Resistance between magnesium and ground rods — 3.75 ohms

* These values may be high if stray currents exist. Otherwise, they will probable not exceed 200 milliamperes.

over-all design. For discussion, these tests will be divided between building ground beds and substation ground installations.

Building Grounds

Four building ground beds were tested during the week of May 14 to determine their ground resistance. These beds had been installed for several months but were not yet connected to the buildings. The results of this test are shown in Table III. Since the design resistance value for these beds had been 3.5 ohms, the designers were very happy with the results. One of these beds was rechecked in May 1953 after it had been connected to the building framework for almost a year, and the resistance was found to be approximately 30 per cent lower than it had

been just after installation. This reduction may be accounted for by the weather, which in the spring of 1953 was exceptionally rainy.

To determine the effect of the magnesium anodes in the ground beds, a detailed exploration of one bed was made by measuring potentials between the various structures involved and a copper-copper sulfate reference electrode placed on the ground surface at several test points throughout the bed. This bed is shown in Fig. 5, with the test points indicated as A, B, C, and D. The measurements were made shortly after the bed had been connected to the building structure; the values obtained are given in Table IV. It is interesting to note that the building framework is definitely cathodic. To obtain a more easily understood picture of this situation, a number of current meas-

urements were made and these are shown diagramatically in Fig. 6. From this diagram, it is quite clear that the building structure is cathodic both to the anodes and to the steel ground rods. It is also quite clear that the anodes are necessary to prevent the ground rods from discharging an appreciable galvanic current into the soil. Since it is reasonably easy to measure these currents, it was decided that this would be a good periodic-maintenance test at each of the ground beds. Potential measurements to copper sulfate electrodes are useful when they can be carefully made and studied, but for general maintenance purposes it is not felt that it would be necessary to use the test cell. Data sheets based on the foregoing conclusions were prepared and in May 1953 these currents were again checked at the same location noted in Fig. 6 and recorded on a prepared data sheet. A data sheet with the readings taken at this time is shown in Table V.

A comparison of the currents measured in 1952 with those measured in 1953 shows that the anodic condition of the steel ground rods has been reduced, and that the magnesium anode current has been increased. All these factors indicate that this particular ground bed is behaving very well. The plant-maintenance personnel plan to make similar measurements on all the ground beds and to repeat them semiannually. In this manner, any significant change will be noted soon enough so that it may be investigated before an unsatisfactory condition has developed too far.
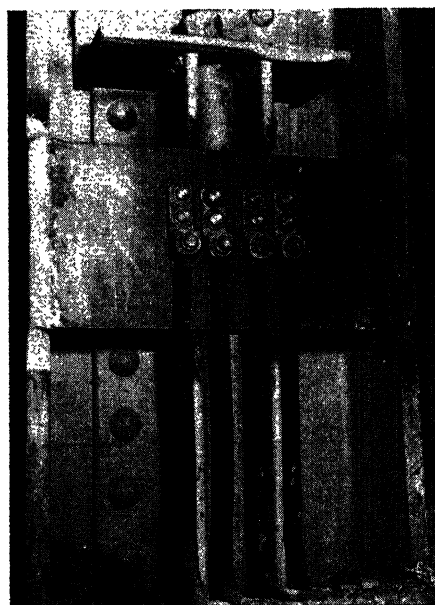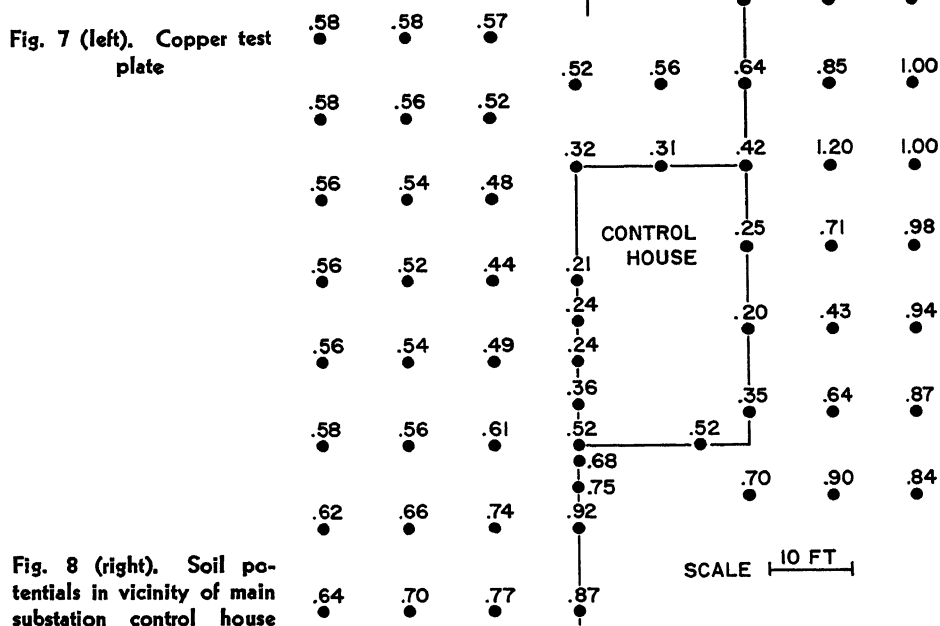
The connection of the magnesium

Fig. 7 (left). Copper test plate

Fig. 8 (right). Soil potentials in vicinity of main substation control house

anodes and the steel ground rods to the building framework is made to a copper plate on the building frame, as shown in Fig. 7. This permits convenient measurement of the quantities under consideration and provides a means of locating a source of stray direct current that may be detected during routine testing at the ground beds. Measurements through 1953 have shown no evidence of stray currents of this nature. But it is to be expected that such currents may be found to exist at some time in the future, and it is important that they be detected and controlled before they can do appreciable damage to the underground system.

Substation Grounds

At the substation locations, it is not possible to isolate the ground connection because such a condition would provide too much of a hazard to safety. However, it is possible to disconnect the anodes from the remainder of the ground network, so that measurements of cathodic protection current and anode resistance can be made. Reference to Fig. 4 will aid in understanding the ground-wiring arrangement at the substation locations.

Tests made in May and June of 1952 were designed to check the effectiveness of the grounding and the cathodic protection at the main substation. Ground-resistance measurements disclosed a value slightly below the design standard of 1 ohm. A potential survey of the area was made by measuring the voltage between a copper-copper sulfate cell (placed at various points) and the ground network. Some of the readings were taken with the magnesium anodes connected and disconnected. The results indicated that the magnesium was, in general, shifting the potential of the steel about 0.1 volt closer to the potential of the magnesium. However, the potential of the soil in the vicinity of the substation control house was found to be very low, a condition that indicated the presence of a large cathodic area. This same condition was found around the concrete foundations near the building ground

beds. In the worst area, it was difficult to raise the potential of the soil by even 0.01 volt with the magnesium. Measurements showed the total anode current to be 1 ampere, a value approximately equivalent to 1 milliampere per square foot of steel ground rod. Although this current density should provide sufficient protection, it was decided to make doubly sure by disconnecting one of the ground rods in the low potential area and measuring the current to that particular rod. The addition of the magnesium reduced the earth-seeking current at this rod from 5.4 milliamperes to 1.0 milliampere. Since this rod was located at one of the lowest potential spots in the area, it was decided that the general condition was quite satisfactory.

These cathodic areas adjacent to the concrete foundations are probably due to the reinforcing steel. It is to be expected that this condition will gradually clear up as the concrete-enclosed steel becomes polarized. The soil potentials taken at this time were carefully made, and were recorded on a map of the area. A portion of this map, reproduced in Fig. 8, shows the potential readings taken in the vicinity of the control house. This figure reveals quite clearly that the cathode must be at or within the control house. Investigation at the site did not reveal any stray direct currents.

Transmission-Tower Grounding

Each transmission-line tower is grounded with four of the 40-foot steel rods, one at each corner of the tower. There are no magnesium anodes for protection of these rods because the latter are generally installed in high-resistance soil areas. However, the concrete structures have proved to be cathodic to the ground rods, so it may eventually be necessary to provide some cathodic protection for the tower grounds. The maintenance force plans to check the resistance to ground of these rods annually as a part of an over-all test program.

Underground Cables and Piping

All underground cables have been neoprene jacketed. Splices have been

wrapped with neoprene tape, and every precaution has been taken to prevent any exposure of the lead or copper to the earth. Under this condition, current can leave the covered metal only at designated drainage points or at incidental holes or breaks in the covering. All buried piping has been coated with coal-tar enamel and then wrapped with coal-tar-saturated asbestos felt. Although the cathodic protection was installed primarily for the steel ground rods, it will provide some measure of protection to the cable sheaths and piping at points where the coating may be damaged by handling during installation.

Stray Direct Currents

Every effort has been made to prevent stray direct currents. It is believed that the planned maintenance and test schedule will provide for their detection before they can do appreciable damage. The ground-bed locations and arrangement are such that tracing stray direct currents will be facilitated. It is recognized that the development of large uncontrolled stray direct currents would offset the benefits of eliminating the bare copper underground, and could cause accelerated corrosion of the magnesium anodes as well as the steel ground rods.

Summary

The over-all underground system at Fairless Works includes four features that tend to aid one another in combating underground corrosion. The elimination of bare copper underground reduces galvanic corrosion; the magnesium anodes protect the steel ground rods and serve to partly protect all underground piping, lead cables, and foundations; the insulating coverings on pipes and cables protect their metallic surfaces and reduce the requirements for cathodic protection; and the design of the ground-bed installations permits analysis of the stray-current situation. The cost considerations and the characteristics of the site have been considered in the design; and it is believed that this installation will serve as a model for many more industrial plants.

Transient Analysis of a D-C Electromagnet With Cutout Switch

T. H. LEE
ASSOCIATE MEMBER AIEE

Synopsis: This paper analyzes the effect of the armature motion on the force stroke characteristics, the pickup speed, and the interruption performance of the cutout switch of a d-c electromagnet. The analysis of the interruption performance is made for a case where a capacitor is connected across the cutout switch. The same method can be extended to other cases.

ONE OF the most important characteristics of a d-c electromagnet is the force-stroke characteristic, frequently called a "pull curve." It indicates the maximum force the electromagnet will develop with a certain excitation and a certain armature gap. It is therefore possible to determine the maximum load from the pull curve and also the maximum gradient of load if spring load is used. This pull curve can be taken by blocking the armature gap and measuring the force exerted by the magnet on the armature after the coil current has reached its steady-state value. Since this curve indicates the force developed with steady-state current flowing in the coil, it will be called the static pull curve to distinguish it from the dynamic pull curve which will be discussed later.

In the case of an a-c electromagnet there is very little difference between the flux density at open armature gap and the sealed-in position of the armature, because the change of reactive voltage at open-gap position, because of the increase in inrush current, is generally a small percentage of the reactive voltage at the sealed-in position. The force of an a-c electromagnet therefore changes slowly as the armature gap is reduced. While in a d-c electromagnet the magnetomotive force is a constant value, the flux is inversely proportional to the reluctance of the magnetic path which in turn varies directly with the air gap. The force, which is proportional to the square of the flux, therefore increases rapidly as the gap is decreased.

To provide sufficient excitation to develop the force required at the open gap position and yet not to overheat the coil if the magnet is to be rated for continuous duty, some coils of d-c electromagnets are wound with two sections. The pickup section is wound with larger wire, and it has a much lower resistance than the holding section. The coils can be connected in one of the three ways shown in Fig. 1. The cutout switch inserts the holding coil in series with the pickup coil in both Fig. 1(A) and (C). In (B) the cutout switch disconnects the pickup coil. On some electromagnets the current in the pickup coil may be so high that the arc will not be extinguished across the cutout switch. Capacitance C connected across the cutout switch is used to help the interruption.[1] The analysis of the interruption performance is limited to the case shown in Fig. 1(A). The same method can be extended to the other two cases. The equivalent circuit for Fig. 1(A) is shown in Fig. 2.

Nomenclature

C = capacitance, farads
E = line voltage, volts
$E_c(t)$ = voltage across capacitance, a function of time
$e_c(s)$ = Laplace transform of $E_c(t)$
F = pull of magnet, pounds
f = a constant for the straight-line equation of pull versus current on log-log paper; a function of gap
g = gravitational constant
h = constant determined from the instantaneous pickup coil current, amperes per second[2]
I = current, amperes
I_1 = current in pickup coil
I_2 = current in holding coil
I_3 = current in capacitance
$i_1(s)$, $i_2(s)$, $i_3(s)$ = Laplace transform of I_1, I_2, I_3
K = peak current of pickup coil during pickup operation
k = spring gradient, pounds per inch
L = self-inductance, henrys
L_1 = self-inductance of pickup coil at opening of cutout switch
L_2 = self-inductance of holding coil at opening of cutout switch
M = mutual inductance between pickup coil and holding coil
m = mass of armature and load
N = number of turns
P = initial spring tension of load, pounds

$$p = -\frac{\beta}{2\alpha c} + \frac{1}{2}\sqrt{\frac{\beta^2}{\alpha^2 c^2} - \frac{4R_2}{\alpha c}}$$

$$q = -\frac{\beta}{2\alpha c} - \frac{1}{2}\sqrt{\frac{\beta^2}{\alpha^2 c^2} - \frac{4R_2}{\alpha c}}$$

R = resistance, ohms

R_1 = resistance of pickup coil
R_2 = resistance of holding coil
s = variable in Laplace transform
t = time, seconds
x = travel of armature, inches
x_n = position of armature at $t = n$ seconds
$\alpha = L_1 R_2 + L_2 R_1$
$\beta = L_1 + L_2 + 2M$
τ = time treated as a constant, seconds
φ = flux, webers

Dynamic Characteristic due to Electromechanical Coupling Between Coil and Armature

The current in the coil at any instant is related to the voltage by the equation

$$E = \frac{d(N\varphi)}{dt} + IR \tag{1}$$

Since N is constant

$$I = \frac{E - N\dfrac{d\varphi}{dt}}{R} \tag{2}$$

The static pull curve is taken with $(d\varphi)/(dt) = 0$. This is evidently not true during the pickup operation since flux does change. In general

$$\varphi = \varphi(I, x) \tag{3}$$

hence

$$\frac{d\varphi}{dt} = \frac{\partial\varphi}{\partial I}\frac{\partial I}{\partial t} + \frac{\partial\varphi}{\partial x}\frac{\partial x}{\partial t} \tag{4}$$

therefore

$$I = \frac{E - N\dfrac{\partial\varphi}{\partial I}\dfrac{\partial I}{\partial t} - N\dfrac{\partial\varphi}{\partial x}\dfrac{\partial x}{\partial t}}{R} \tag{5}$$

If $x = 0$ is chosen at the open-gap position, then $(\partial\varphi)/(\partial x)$, $(\partial x)/(\partial t)$, and $(\partial\varphi)/(\partial I)$ are all positive. When the armature is not moving, $(\partial x)/(\partial t) = 0$ and the equation reduces to

$$I = \frac{E - N\dfrac{\partial\varphi}{\partial I}\dfrac{dI}{\partial t}}{R} = \frac{E - L\dfrac{\partial I}{\partial t}}{R} \tag{6}$$

As soon as the armature starts to move, the term $\{N[(\partial\varphi)/(\partial x)][(\partial x)/(\partial t)]\}$ is always positive and therefore tends to reduce the instantaneous current. The term $(\partial\varphi)/(\partial x)$ actually is proportional to the variation of permeability with respect to the armature gap. As the gap becomes smaller, $(\partial\varphi)/(\partial x)$ will become greater. As the armature travels further toward the sealed in position, both terms $(\partial\varphi)/(\partial x)$ and $(\partial x)/(\partial t)$ increase and
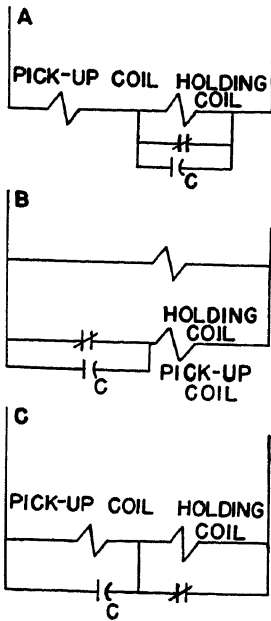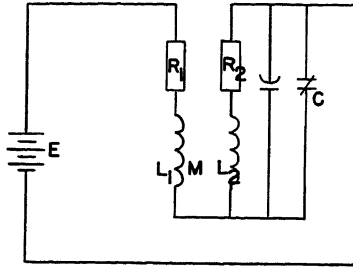
Fig. 1 (left). Connection of 2-section coil of a d-c electromagnet with cutout switch and capacitor

Fig. 2 (right). Equivalent circuit for Fig. 1(A)

therefore tend to reduce the current more. The instantaneous pull at any gap varies as the square of the instantaneous current. This reduction in current due to the mechanical motion of the armature thus affects the dynamic characteristics of the electromagnet greatly.

Roters[2] illustrates a step-by-step method of calculating the dynamic characteristics. This method, tedious as it is, is of great importance to designers of a d-c electromagnet with cutout switch because the dynamic characteristics affect directly the interruption performance of the cutout switch. At the end of this paper a simpler step-by-step method with the help of a magnetic oscillograph will be illustrated.

Analysis of Interruption Performance of Cutout Switch

The interruption of direct current with an ordinary switch is much more difficult than the interruption of alternating current because the switch must actually force the arc to become unstable.[3] With a capacitance across the switch the situation is entirely different. At the instant the contacts of the switch are separated, the voltage across the contacts is zero because the voltage across the capacitance cannot change instantly. A considerable recovery strength is developed almost immediately across the switch because of the formation of a sheath of neutral gas around the contacts. As the movable contact of the switch travels farther, the recovery strength increases with the rate of increase depending on the speed of the contact. A curve of recovery strength versus time will have the general shape

shown in Fig. 3. In the meantime, the capacitance charges up. If the voltage across the capacitance is at no time higher than the recovery strength, there will be no arc across the switch and the energy stored in the pickup coil will be dissipated in the resistances. If the voltage across the capacitance is higher than the recovery strength, ignition will occur between the contacts. As soon as ignition starts, the capacitance will discharge rapidly through the arc and the voltage across the switch will drop to the value corresponding to the current and gap. If the switch is not able to force the arc to become unstable, the arc will not extinguish. The capacitance across the switch therefore changes the principle of interruption from forcing the arc to become unstable to preventing the ignition across the contacts.

The analysis of interruption performance of the cutout switch consists therefore of two problems:

1. Determining the recovery strength of the cutout switch.

2. Calculating the voltage across the cutout switch.

The recovery strength of the switch with a capacitance in parallel is hard to determine. It must have a higher value than the recovery strength of the same switch at the natural current zero of a-c interruption because it does not have a previously ionized column which exists between the contacts during a-c interruption. Experiments on some electromagnets with capacitance not high enough to extinguish the arc indicated that the recovery strength is less than the natural spark breakdown strength. For one particular switch, the upper limit of its recovery strength can be estimated by an experiment.

Connect up a d-c electromagnet as shown in Fig. 1(A). With the armature blocked in the sealed-in position and line voltage applied, open the cutout switch with a similar magnet. Start with a value of C low enough that the cutout switch cannot interrupt the current. Then increase the value of C until the cutout switch interrupts satisfactorily. Using this value of C and equation 11,

which will be derived later, it is possible to calculate the curve of voltage across the cutout switch versus time. This curve can be used as the upper limit of the recovery strength. It is true that the recovery strength depends upon the opening speed of the contacts, but if the cutout switch is opened at the lowest operating speed in practical appplication the calculated upper limit of the recovery strength will be on the safe side.

The electromagnet shown in Fig. 1(A) can be represented by the equivalent circuit shown in Fig. 2. Since both coils are wound on the same core, it is fairly safe to assume that the coupling between two coils is perfect, or

$$M = \sqrt{L_1 L_2} \qquad (7)$$

In calculating the voltage across the cutout switch, the question of what is the boundary condition of currents in the two coils immediately presents itself. The most severe condition, as far as interruption by the cutout switch is concerned, generally occurs when the current in the pickup coil is equal to the steady-state value. This corresponds to a load curve which coincides with the static pull curve. The voltage across the cutout switch will be derived for this case since the resulting equation is useful in estimating the upper limit of recovery strength of the cutout switch. The effect of armature motion on the voltage across the cutout switch will be investigated later.

The equations for the circuit in Fig. 2 are

$$E = R_1 I_1 + L_1 \frac{dI_1}{dt} + M \frac{dI_2}{dt} + R_2 I_2 + L_2 \frac{dI_2}{dt} + M \frac{dI_1}{dt} \qquad (8)$$

$$\int \frac{I_3 dt}{C} = R_2 I_2 + L_2 \frac{dI_2}{dt} + M \frac{dI_1}{dt} \qquad (9)$$

$$I_3 = I_1 - I_2 \qquad (10)$$

At $t = 0$ (when the cutout switch opens)

$$I_1(0) = \frac{E}{R_1}, \quad I_2(0) = 0$$

The solution of these equations (Appendix I) gives the voltage across the cutout switch as

$$E_c(t) = E\left[1 + \left(\frac{1}{R_1 C(p-q)} + \frac{q}{p-q}\right)\epsilon^{pt} - \left(\frac{1}{R_1 C(p-q)} + \frac{p}{p-q}\right)\epsilon^{qt}\right] \qquad (11)$$

where

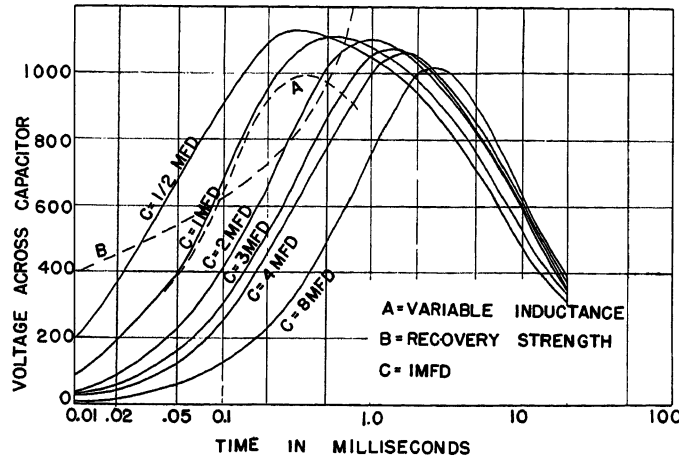$$p = \frac{-\beta}{2\alpha C} + \frac{1}{2}\sqrt{\frac{\beta^2}{\alpha^2 C^2} - \frac{4R_2}{\alpha C}}$$

Fig. 3. Voltage across cutout switch versus time

Fig. 4 (above). (A) 60-cycle timing curve. (B) Voltage across cutout switch. (C) Pickup coil current

$$q=\frac{-\beta}{2\alpha C}-\frac{1}{2}\sqrt{\frac{\beta^2}{\alpha^2 C^2}-\frac{4R_2}{\alpha C}}$$

$$\alpha=L_2R_1+L_1R_2$$

$$\beta=L_1+L_2+2M$$

In arriving at equation 11, the following approximations were used

$$R_1 << R_2$$

$$R_1R_2C << L_1+L_2+2M$$

A sample calculation has been made and the results are plotted in Fig. 3 for various values of C and the following parameters:

$L_1=1$ henry $R_1=25$ ohms
$L_2=16$ henrys $R_2=2,500$ ohms
$M=4$ henrys

In actual applications, the load curve will always lie under the static pull curve at 90 or 85 per cent of the rated voltage so that the magnet will surely pick up. The current in the pickup coil at the instant of the opening of the cutout switch is always less that the steady-state value. Also, the current in the holding coil at that instant is not zero due to the coupling between the pickup and the holding coils. The voltage across the cutout switch could be considerably different from the results calculated with equation 11.

A typical oscillogram of pickup coil current is shown in Fig. 4. The current has a negative slope just before the cutout switch opens. It is necessary to determine the boundary conditions for equations 8 through 10 for calculating the voltage across the cutout switch. The descending portion of pickup coil current can be closely approximated by the parabola of equation 12

$$I_1(t^1)=K-ht'^2 \tag{12}$$

where t' has the value zero when the pickup coil current is at its peak, K is the

peak of pickup coil current, and h is a constant depending on the speed of pickup. The higher the speed of pickup, the higher will be the value of h.

The value of I_2 can be found by substituting equation 12 into and solving 13

$$R_2I_2+L_2\frac{dI_2}{dt^1}+M\frac{dI_1}{dt^1}=0 \tag{13}$$

The parameters L_1, L_2, and M vary between the time $t'=0$ and the time of opening of the cutout switch. If the positions of the armature at these two instants differ greatly, it will be necessary to move the point $t'=0$ to some point closer to the opening of the cutout switch so that the values of L_1, L_2, and M at the armature position corresponding to the opening of the cutout switch can be used. For the calculations in this paper it will be assumed that the positions of the armature corresponding to these two instants are very close together and values of L_1, L_2, and M will be those at the opening of the cutout switch.

Solution of equation 13 is

$$I_2(t')=\frac{2hM}{L_2}\left[\frac{L_2^2}{R_2^2}\left(\epsilon^{-\frac{R_2}{L_2}t'}-1\right)+\frac{L_2}{R_2}t'\right] \tag{14}$$

Expanding the term $\epsilon^{-\frac{R_2}{L_2}t'}$ into a series, using

$$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\dots \tag{15}$$

and considering that t is of the order of milliseconds so that terms higher than t^3 can be neglected, then

$$I_2(t')=\frac{2hM}{L_2}\left(\frac{1}{2}t'^2-\frac{1}{6}\frac{R_2}{L_2}t'^3\right) \tag{16}$$

Let the cutout switch open at $t'=\tau$; then the boundary conditions for solving equations 8, 9, and 10 become

$$I_1(0)=K-h\tau^2 \tag{17}$$

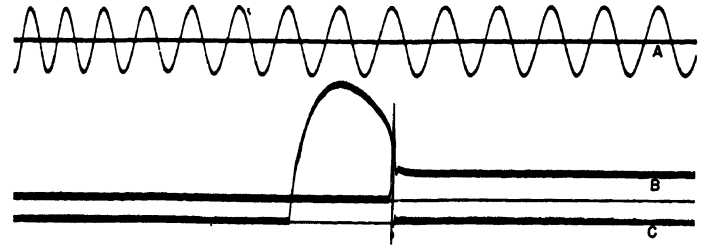$$I_2(0)=\frac{2hM}{L_2}\left(\frac{1}{2}\tau^2-\frac{1}{6}\frac{R_2}{L_2}\tau^3\right) \tag{18}$$

Assuming that I_1 reaches its steady-state value before it starts to decrease (this corresponds to a loading coinciding with the static pull curve for a certain portion of armature travel), the solution of the voltage across the cutout switch $E_c(t)$ with the help of the Laplace transformation is then (see Appendix II)

$$E_c(t)=\frac{E}{R_1C}\left[1+\frac{hMR_2R_1}{3\alpha L_2E}\tau^3\left(R_1-\frac{MR_2}{L_2}\right)\right]\times$$
$$\frac{1}{p-q}(\epsilon^{pt}-\epsilon^{qt})+\frac{ER_2}{\alpha C}\frac{1}{p-q}\left(\frac{1}{p}\epsilon^{pt}-\frac{1}{q}\epsilon^{qt}\right)+E \tag{19}$$

In equation 19 only the coefficient of the first term on the right-hand side depends on the boundary conditions. In many electromagnets the second term

$$\frac{ER_2}{\alpha c}\frac{1}{p-q}\left(\frac{1}{p}\epsilon^{pt}-\frac{1}{q}\epsilon^{qt}\right)$$

is much smaller than the first term and therefore the quantity

$$r=1+\frac{hMR_2R_1}{3\alpha L_2E}\tau^3\left(R_1-\frac{MR_2}{L_2}\right) \tag{20}$$

is of interest.

The holding coil of an electromagnet is always wound with smaller wire then the pickup coil and hence generally

$$\frac{R_1}{L_1}<\frac{R_2}{L_2}\quad\text{or}\quad R_1<\frac{L_1R_2}{L_2}$$

Since L_2 is larger then L_1, M is also larger than L_1. Therefore

$$R_1<\frac{MR_2}{L_2}$$

r is therefore generally less than 1. When the second term in the right-hand side of equation 19 is negligible compared to the first term, the value of r is an indication of the reduction in voltage across the cutout switch due to the armature motion. The reduction in voltage can affect the selection of capacitance in two ways:

1. A much smaller capacitance is sufficient for interruption.

2. A capacitance of lower voltage rating can be used due to the reduction of peak voltage.

r can be written in terms of $I_1(0)$ as

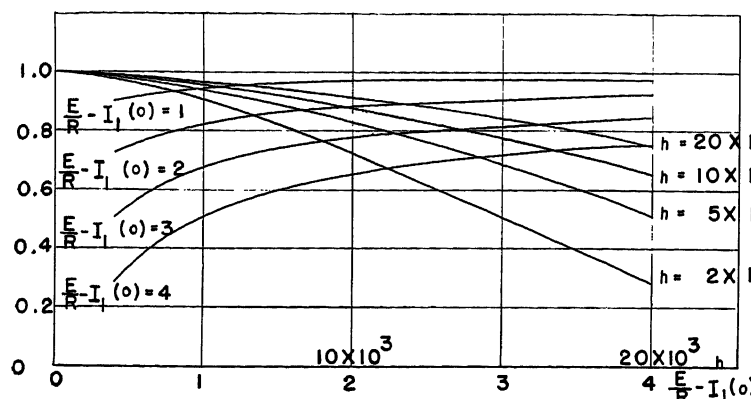$$r=1+u\sqrt{\frac{1}{h}\left(\frac{E}{R_1}-I_1(0)\right)^3} \tag{21}$$

Fig. 5. Value of r for constant h and for constant E/R-I₁(0)

Fig. 6 (right). Displacement versus time

where

$$u = \frac{MR_1R_2}{3\alpha EL_2}\left(R_1 - \frac{MR_2}{L_2}\right)$$

h is an indication of the speed of pickup. The higher the value of h, the higher is the speed of pickup. Speed of pickup depends on load. A dead weight load generally has a lower speed of pickup than a spring load of the same amount. The term $[E/R_1 - I_1(0)]$ is an indication of the position where the cutout switch opens for the same load, and consequently the same h. Families of curves plotted with r versus $[E/R_1 - I_1(0)]$ for constant h and with r versus h for constant $I_1(0)$ are shown in Fig. 5.

It is interesting to note that, for constant $I_1(0)$, the higher the speed of pickup the higher the voltage across the cutout switch and hence the more difficult is the interruption. For constant h, the lower the value $I_1(0)$ the easier the interruption, as expected. The curves in Fig. 5 are not only useful in understanding the interruption performance but can also be used quantitatively in estimating the reduction of voltage across the cutout switch. For any electromagnet, the highest possible voltage across the cutout switch can be estimated if the worst loading condition is known because the load determines the value of h. The position where the cutout switch is set to open determines the value $I_1(0)$.

Calculation of Displacement Versus Time and Dynamic Pull Curve from Oscillogram of Current Versus Time

If the static pull curve and the instantaneous current for a particular magnet and given load are known from experiment, the instantaneous position of the armature and the corresponding instantaneous pull can be calculated by the method described now. In the following

derivation, hysteresis and eddy currents are neglected. Test results have justified this simplification. There is also negligible saturation effect before the armature is sealed.

With no saturation effect, the static pull curve, plotted as force versus current for a certain gap is a straight line with a slope of 2 on log-log paper since force is proportional to the square of the flux and flux is directly proportional to current if there is no saturation. The pull can therefore be represented by

$$\log F = \log I^2 + \log f(x) \quad (22)$$

or

$$F = f(x)I^2 \quad (23)$$

where $f(x)$ is a function of gap.

For any current I_0

$$F(I_0, x) = I_0^2 f(x) \quad (24)$$

where $F(I_0,x)$ is the static pull curve for current I_0, or

$$f(x) = \frac{F(I_0, x)}{I_0^2} \quad (25)$$

The instantaneous pull is a function of time since current is a function of time.

$$F(t) = \frac{F(I_0,x)}{I_0^2} I^2(t) \quad (26)$$

To find the instantaneous pull, it is necessary to know the position of armature at different times. The equation of motion of the armature is

$$m\frac{d^2x}{dt^2} = F(t) - mg - (P + kx) \quad (27)$$

Substitute

$$\frac{d^2x}{dt^2} = \frac{x_{n+1} - 2x_n + x_{n-1}}{(\Delta t)^2} \quad (28)$$

into equation 27

$$m\frac{x_{n+1} - 2x_n + x_{n-1}}{(\Delta t)^2} = \frac{I^2(t_n)}{I_0^2}F(I_0 x_n) - mg - P - kx_n \quad (29)$$
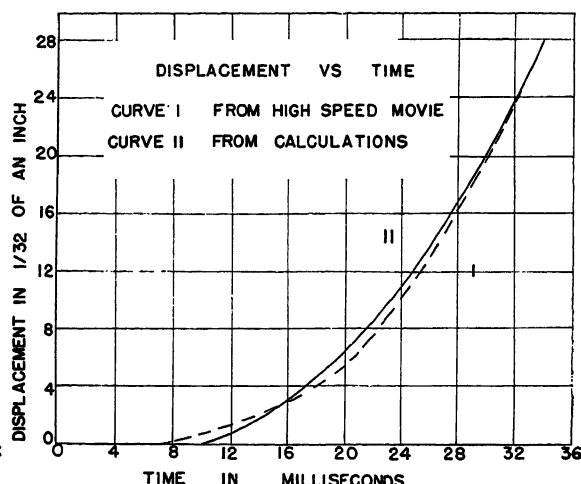
Equation 29 can be solved numerically

by using the static pull curve and the oscillogram of the instantaneous current. Fig. 6 compares the calculated curve for one sample and the experimental curve from a high-speed movie. The first term on the right-hand side of equation 29 gives the instantaneous pull for $t = t_n$.

Remarks

In calculating the transient voltage across the cutout switch, the inductances of the pickup and holding coils are assumed to be constant. An error is obviously introduced since the inductances increase as the armature moves toward the sealed-in position. To treat this condition, the inductances have to be considered as a function of time and the differential equations thus have variable coefficients. It is fairly accurate to assume that the inductance varies inversely as the armature gap for a small range of gap such as from the position where the cutout switch opens to the sealed-in position. The speed of the armature at this range is also fairly constant since the speed is relatively high and the distance is small. The inductances can then be represented by $L_1(t) = L_1(1+kt)$, $L_2(t) = L_2(1+kt)$, and $M(t) = M(1+kt)$, where k can be determined by combining the variation of inductances and speed of the armature within the required range. The solution of the equations can be obtained by assuming two infinite series around $t = 0$ for both I_1 and I_2 as shown in Appendix III. A calculation has been made for a value of $k = 250$, $c = 1$ microfarad, and other parameters the same as before. It is plotted in Fig. 3 and lies under the voltage curve for the same value of C and constant inductances. The energy stored in the pickup coil at the instant the cutout switch opens is $1/2L_1I_1^2$. As L_1 increases with time, I_1 decreases since the stored energy in the pickup coil cannot increase any more after the open-

Lee—Transient Analysis of D-C Electromagnet With Cutout Switch

ing of the cutout switch. Therefore, the pickup coil will actually try to maintain a current less than the value for a constant inductance. This decrease in current is directly responsible for the lower voltage across the cutout switch.

The calculation by assuming an infinite series is long and tedious. Since the results obtained by assuming constant inductance are on the safe side, the variation of inductance can be neglected.

The dynamic characteristic not only affects the interruption performance of the cutout switch but also directly affects performance data such as pickup time, duty-cycle rating, and mechanical life. For a pure mechanical system, an increase of force by 21 per cent reduces the time required to travel the same distance by 10 per cent. Because of the interrelation between the electric circuit and mechanical motion, an increase of 10 per cent in voltage will not cut down the pickup time by 10 per cent. The duty-cycle rating of the electromagnet can be defined only for a particular load since the magnitude and duration of current depend on load. The mechanical energy stored in the armature is the area enclosed by the dynamic pull curve and the axis corresponding to zero pull. The impact on the seat and hence mechanical life is dependent also on the dynamic characteristic. Unfortunately, a simple analysis is impossible because of the complicated nonlinear relationship of the magnetic circuits.

Appendix I. Solution of Equations 8, 9, and 10

Solving equation 10 for I_2 and substituting in equations 8 and 9

$$E = (L_1+L_2+2M)\frac{dI_1}{dt} - (L_2+M)\frac{dI_3}{dt} + (R_1+R_2)I_1 - R_2I_3 \quad (30)$$

$$\int \frac{I_3 dt}{c} = (L_2+M)\frac{dI_1}{dt} + R_2I_1 - L_2\frac{dI_3}{dt} - R_2I_3 \quad (31)$$

Using the Laplace transformation[4] with the boundary conditions $I_1(0) = I_3(0) = E/R_1$, the following equations are obtained

$$[(L_1+L_2+2M)s+R_1+R_2]i_1(s) - [(L_2+M)s+R_2]i_3(s) = \frac{E}{s} + \frac{E}{R_1}(L_1+M)$$
$$= E\left(\frac{1}{s} + \frac{L_1+M}{R_1}\right) \quad (32)$$

$$[(L_2+M)s+R_2]i_1(s) - \left(L_2s+R_2+\frac{1}{Cs}\right)i_3(s) = M\frac{E}{R_1} \quad (33)$$

Solving for $i_3(s)$

$$i_3(s) = \frac{\begin{vmatrix} (L_1+L_2+2M)s+R_1+R_2 & E\left(\frac{1}{s}+\frac{L_1+M}{R_1}\right) \\ (L_2+M)s+R_2 & E\frac{M}{R_1} \end{vmatrix}}{\begin{vmatrix} (L_1+L_2+2M)s+R_1+R_2 & -(L_2+M)s-R_2 \\ (L_2+M)s+R_2 & -L_2s-R_2-\frac{1}{Cs} \end{vmatrix}} \quad (34)$$

$$i_3(s) = E \frac{s\left(\frac{M}{R_1}(L_1+L_2+2M) - \frac{1}{R_1}(L_1L_2+M^2+L_1M+L_2M)\right) - \left((L_2+M)+\frac{R_2}{R_1}L_1-M-\frac{R_2}{s}\right)}{-\frac{1}{Cs}[(R_1L_2+R_2L_1)Cs^2+(R_1R_2C+L_1+L_2+2M)s+R_1+R_2]}$$

Since perfect coupling is assumed, the coefficient of s in the numerator vanishes as does the coefficient of s^2 in the denominator. It is a good approximation for most cases to assume that R_1R_2C is negligible compared to L_1+L_2+2M and R_1 is negligible compared to R_2.

Let

$$\alpha = R_1L_2+R_2L_1 \quad (35)$$

$$\beta = L_1+L_2+2M \quad (36)$$

Then

$$i_3(s) = E \frac{-\left(L_2+\frac{R_2}{R_1}L_1\right)-\frac{R_2}{s}}{-\frac{1}{Cs}(\alpha Cs^2+\beta s+R_2)}$$

$$= \frac{\frac{E}{R_1}\left((L_2R_1+R_2L_1)+\frac{R_1R_2}{s}\right)}{\frac{1}{Cs}(\alpha Cs^2+\beta s+R_2)}$$

$$= \frac{E}{R_1}\frac{Cs\left(\alpha+\frac{R_1R_2}{s}\right)}{\alpha Cs^2+\beta s+R_2} \quad (37)$$

Since

$$E_c(t) = \int_0^t \frac{I_3 dt}{C} \quad (38)$$

then

$$e_c(s) = \frac{1}{Cs}i_3(s) \quad (39)$$

$$e_c(s) = \frac{E}{R_1}\left(\frac{\alpha}{\alpha Cs^2+\beta s+R_2} + \frac{R_1R_2}{s(\alpha Cs^2+\beta s+R_2)}\right) = \frac{E}{R_1C} \times$$
$$\frac{1}{s^2+\frac{B}{\alpha C}s+\frac{R_2}{\alpha C}} + \frac{ER_2}{\alpha C} \times \frac{1}{s\left(s^2+\frac{\beta s}{\alpha C}+\frac{R_2}{\alpha C}\right)} \quad (40)$$

Let

$$p = -\frac{\beta}{2\alpha C} + \frac{1}{2}\sqrt{\frac{\beta^2}{\alpha^2C^2} - \frac{4R_2}{\alpha C}} \quad (41)$$

$$q = -\frac{\beta}{2\alpha C} - \frac{1}{2}\sqrt{\frac{\beta^2}{\alpha^2C^2} - \frac{4R_2}{\alpha C}} \quad (42)$$

then

$$e_c(s) = \frac{E}{R_1C}\frac{1}{p-q}\left(\frac{1}{s-p} - \frac{1}{s-q}\right) + \frac{ER_2}{\alpha C}\left[\frac{1}{p-q}\left(\frac{1}{p(s-p)} - \frac{1}{q(s-q)}\right)\right] + \frac{E}{s} \quad (43)$$

and by taking inverse transform,

$$\frac{E_c(t)}{E} = 1 + \frac{1}{p-q}\left[\left(\frac{1}{R_1C}+q\right)\epsilon^{pt} - \left(\frac{1}{R_2C}+p\right)\epsilon^{qt}\right] \quad (44)$$

Appendix II. Solution of Equations 8, 9, and 10 with Initial Conditions Given by Equations 17 and 18

The Laplace transform of equation 8 with boundary conditions as shown in equations 17 and 18 is

$$\frac{E}{s} + \beta\left(\frac{E}{R_1} - h\tau^2\right) - (L_2+M)\left(\frac{E}{R_1} - h\tau^2 - \frac{hM}{L_2}\tau^2 + \frac{hMR_2}{3L_2^2}\tau^3\right) = \beta s i_1(s) - (L_2+M)s i_3(s) + (R_1+R_2)i_1(s) - R_2i_3(s) \quad (45)$$

Since

$$(L_2+M)\left(1+\frac{M}{L_2}\right) = \frac{L_2^2+2L_2M+M^2}{L_2} = L_2+L_1+2M = \beta$$

this equation becomes

$$\frac{E}{s} + (L_1+M)\frac{E}{R_1} - \left(1+\frac{M}{L_2}\right)\frac{hMR_2}{3L_2}\tau^3 = (\beta s+R_2+R_1)i_1(s) - [(L_2+M)s+R_2]i_3(s) \quad (46)$$

Similarly, the transform of equation 9 is

$$\left(\frac{ME}{R_1} - \frac{hMR_2}{3L_2}\tau^3\right) = [(L_2+M)s+R_2]i_1(s) - \left(L_2s+R_2+\frac{1}{Cs}\right)i_3(s) \quad (47)$$

The solution $i_3(s)$ will have the same determinant in the denominator as equation 34 and an additional part in the numerator, which is

$$\begin{vmatrix} \beta s + R_1 + R_2 & -(L_2+M)\dfrac{hMR_2}{3L_2^2}\tau^3 \\ (L_2+M)s+R_2 & -\dfrac{hMR_2}{3L_2}\tau^3 \end{vmatrix} = \frac{-hMR_2}{3L_2}\tau^3\begin{vmatrix} \beta s + R_1 + R_2 & 1+\dfrac{M}{L_2} \\ (L_2+M)s+R_2 & 1 \end{vmatrix}$$

$$=\frac{-hMR_2}{3L_2}\tau^3\left(\beta s+R_1+R_2-(L_2+M)s-R_2-\frac{M}{L_2}(L_2+M)s-\frac{MR_2}{L_2}\right)$$

$$=\frac{-hMR_2}{3L_2}\tau^3\left(R_1-\frac{MR_2}{L_2}\right) \quad (48)$$

Hence

$$i_3(s)=Cs\frac{\dfrac{E}{R_1}\left[\alpha+\dfrac{hMR_2R_1}{3L_2E}\tau^3\left(R_1-\dfrac{MR_2}{L_2}\right)\right]+\dfrac{ER_2}{s}}{\alpha Cs^2+\beta s+R_2} \quad (49)$$

Therefore

$$e_o(s)=\frac{E}{R_1}\left[\alpha+\frac{hMR_2R_1}{3L_2E}\tau^3\left(R_1-\frac{MR_2}{L_2}\right)\right]\frac{1}{\alpha Cs^2+\beta s+R_2}+\frac{ER_2}{s(\alpha Cs^2+\beta s+R_2)}$$

$$=\frac{E}{R_1C}\left[1+\frac{hMR_2R_1}{3\alpha L_2E}\tau^3\left(R_1-\frac{MR_2}{L_2}\right)\right]\frac{1}{s^2+\dfrac{\beta}{\alpha C}s+\dfrac{R_2}{\alpha C}}+\frac{ER_2}{\alpha C}\frac{1}{s\left(s^2+\dfrac{\beta}{\alpha C}s+\dfrac{R_2}{\alpha C}\right)} \quad (50)$$

$$E_o(t)=\frac{E}{R_1C}\left[1+\frac{hMR_2R_1}{3\alpha L_2E}\tau^3\left(R_1-\frac{MR_2}{L_2}\right)\right]\frac{1}{p-q}[\epsilon^{pt}-\epsilon^{qt}]+\frac{ER_2}{\alpha C}\frac{1}{p-q}\left[\frac{1}{p}\epsilon^{pt}-\frac{1}{q}\epsilon^{qt}\right]+E \quad (51)$$

Appendix III. Solution of Equations 8, 9, and 10 with Variable Inductances

If $L_1(t)=L_1(1+kt)$, $L_2(t)=L_2(1+kt)$, and $M(t)=M(1+kt)$, then equations 8 and 9 become

$$E=R_1I_1+\frac{d}{dt}[L_1(t)I_1]+\frac{d}{dt}[M(t)I_2]+R_2I_2+\frac{d}{dt}[L_2(t)I_2]+\frac{d}{dt}[M(t)I_1]$$

$$=R_1I_1+L_1\frac{d}{dt}[I_1(1+kt)]+M\frac{d}{dt}[I_2(1+kt)]+R_2I_2+L_2\frac{d}{dt}[I_2(1+kt)]+M\frac{d}{dt}[I_1(1+kt)] \quad (52)$$

$$\int\frac{I_2dt}{C}=R_2I_2+\frac{d}{dt}[L_2(t)I_2]+\frac{d}{dt}[M(t)I_1]$$

$$=R_2I_2+L_2\frac{d}{dt}[I_2(1+kt)]+M\frac{d}{dt}[I_1(1+kt)] \quad (53)$$

Assuming a series around $t=0$

$$I_1=a_0+a_1t+a_2t^2+\ldots+a_nt^n+\ldots \quad (54)$$

and

$$I_3=b_0+b_1t+b_2t^2+\ldots+b_nt^n+\ldots \quad (55)$$

where $a_0=b_0=I_1(0)$. Substituting equations 54 and 55 into 52 and 53, two simultaneous recursion equations are obtained for the coefficients

$$[R_1+R_2+\beta k(n+1)]a_n-[R_2+(L_2+M)\times(n+1)k]b_n+\beta(n+1)a_{n+1}-(L_2+M)(n+1)b_{n+1}=0 \quad (56)$$

when

$n=0$

$$\beta a_1-(L_2+M)b_1=E-R_1I_1(0)-(L_1+M)kI_1(0)$$

$$\frac{b_{n-1}}{C_n}=[R_2+(L_2+M)(n+1)k]a_n-[R_2+L_2k(n+1)]b_n+(L_2+M)\times(n+1)a_{n+1}-L_2(n+1)b_{n+1} \quad (57)$$

References

1. MAGNETIC CONTROL OF INDUSTRIAL MOTORS (book), G. W. Heumann. John Wiley & Sons, New York, N. Y., 1947, pp. 156–58.

2. ELECTROMAGNETIC DEVICES (book), H. C, Roters. John Wiley & Sons, New York, N. Y., 1st ed., 1941, pp. 373–91.

3. GASEOUS CONDUCTORS (book), J. D. Cobine. McGraw Hill Book Company, New York, N. Y., 1941, pp. 348–63, 371–82.

4. MODERN OPERATIONAL MATHEMATICS IN ENGINEERING (book), R. V. Churchill. McGraw-Hill Book Company, New York, N. Y., 1944.

Transient Analysis of A-C Servomechanisms

S. S. L. CHANG
ASSOCIATE MEMBER AIEE

THIS paper deals essentially with two aspects of the a-c servo problem:

1. It introduces a method whereby the transient data transfer functions are derived directly from the d-c transfer functions.

2. It presents a method of taking into account the electrical transients in rotating components. These electrical transient effects are important to performance as well as stability in the upper data frequency range.

A position control a-c servo with carrier frequency and phase errors is analyzed to illustrate the methods.

Much valuable work relating to these subjects can be found in the literature, and it is desirable to state here what further results can be accomplished with this paper. The complex transform method enables one to calculate the transient response directly and it is also numerically simpler than Sobczyk's method[1] of upper and lower side bands in analyzing a sinusoidal input. Its numerical simplicity is illustrated by calculating the carrier frequency and phase shift problem. Sobczyk (and later Attura[2]) assumed the presence of either frequency error or phase error but not both and came to the conclusion that the performance of the system is not seriously affected. Both the experimental evidence and this method show that with both errors present the system performance can be very detrimentally affected according to a manner predicted by the theory.

Regarding to the transient analysis of rotating components, Brown[3] took

into account the effect of rotor leakage reactance, but neglected the stator leakage reactance and the amplifier impedance in his derivation of transfer function for a 2-phase servomotor. Frazier[4] derived a characteristic equation for electrical transients in an induction tachometer under rather restricted terminating conditions. By the method of variations, the transient equivalent circuits derived in this paper take into account all motor constants as well as terminating network elements. The results are simple enough to apply.

Complex Data Transfer Function

The instantaneous value of a carrier signal can be represented as

$$\sigma(t) = e_1(t) \sin \omega_c t + e_2(t) \cos \omega_c t \qquad (1)$$

Generally both the sine and cosine terms are present, and the transmitted data is contained in the envelope functions $e_1(t)$ and $e_2(t)$ only.

If $e_1(t)$ and $e_2(t)$ are assumed to be arbitrary functions without restriction, they are not uniquely defined by equation 1. That is, there are other sets of $e_1(t)$ and $e_2(t)$ which give the same instantaneous signal $\sigma(t)$. For example: $e_1(t) + \cos \omega_c t$ and $e_2(t) - \sin \omega_c t$ constitute such a set. However, if $e_1(t)$ and $e_2(t)$ are restricted to slow varying functions without appreciable frequency components of the order of carrier frequency or higher, then their values are unique, and their property of slow varying is preserved in passing through a linear network. Physically, the condition of slow varying is justifiable in an a-c servomechanism as the transmitted data are invariably of lower frequency than the carrier.

The partial phase-shift operator j is defined to operate on the carrier only by advancing its phase 90 degrees, while leaving the envelope functions unaffected. Symbolically, equation 1 may be written as

$$\sigma(t) = [e_1(t) + je_2(t)] \sin \omega_c t \qquad (2)$$

Since the carrier factor $\sin \omega_c t$ is common to all signals, it can be left out of the expression altogether. The entire signal is represented by a complex envelope function

$$e(t) = e_1(t) + je_2(t) \qquad (3)$$

In the special case of both e_1 and e_2 being constants, equation 3 reduces to the vector expression of alternating current.

In passing through a linear network, the output envelope $e_o(t)$ and input

envelope $e_i(t)$ are related by the following equation

$$e_o(t) = G(p_e)e_i(t) \qquad (4)$$

$G(p_e)$ is a complex function of the envelope time derivative operator p_e. Let $H(p)$ represent the operator which relates the instantaneous output signal $\sigma_o(t)$ and input signal σ_i such that

$$\sigma_o(t) = H(p)\sigma_i(t) \qquad (5)$$

where p is the instantaneous time derivative operator. The operator $G(p_e)$ can be derived from $H(p)$ by the following substitution

$$G(p_e) = H(p_e + j\omega_c) \qquad (6)$$

In equations 4 and 5, $G(p_e)$ operates on the envelope $e_i(t)$ only, while $H(p)$ operates on the full expression of the input signal including both envelopes and carrier factors. The time constants in $H(p)$ do not have a direct bearing on data transmission, while the time constants in $G(p_e)$ represent a delay or an advance in phase of the transmitted data. The envelope operator $G(p_e)$ in transient analysis is the counterpart of the data-frequency transfer ratio in sinusoidal analysis.

The operators p_e and j in $G(p_e)$ can be manipulated algebraically as a real algebraic number and $\sqrt{-1}$ respectively. It is also permissible to rationalize the expression $G(p_e)$ or to expand it in a power series of p_e.

To show that $G(p_e)$ reduces to the data-frequency transfer ratio in the case of a sinusoidal modulating signal, let $G_d(p_e)$ and $G_q(p_e)$ denote the real and imaginary components of $G(p)$. In equation 4, the real and imaginary components of e_i and e_o are the envelopes of the $\sin \omega_c t$ and $\cos \omega_c t$ terms respectively. In the literature of sinusoidal analysis, they are referred to as direct and quadrature components. Accordingly, G_d transfers a direct component into a direct component, a quadrature component into a quadrature component; G_q transfers a direct component into a quadrature component and vice versa. It remains to be shown that G_d and G_q reduce to their corresponding expressions in sinusoidal analysis when the envelopes are sinusoidal.

By definition, from equation 6

$$G_d(p_e) = \frac{1}{2}[H(p_e + j\omega_c) + H(p_e - j\omega_c)] \qquad (7)$$

In equation 7, as j terms are cancelled in the total expression, its operational significance is immaterial. Its only significance is being $\sqrt{-1}$, and it can be substituted by any other operator $\sqrt{-1}$.

Let ω_m be the data frequency, and i be the data phase-shift operator. Operating on a sinusoidal envelope, $G_d(p_e)$ becomes

$$G_d(i\omega_m) = \frac{1}{2} \ [H[i(\omega_m + \omega_c)] + H[i(\omega_m - \omega_c)]]$$

$$= -(H^+ + H^{*-}) \qquad (8)$$

Similarly

$$G_q(i\omega_m) = \frac{1}{2i}[H[i(\omega_m + \omega_c)] - H[i(\omega_m - \omega_c)]]$$

$$= \frac{1}{2i}(H^+ - H^{*-}) \qquad (9)$$

Equations 8 and 9 are to be compared with equations 1 and 2 in Bjornson's paper.[5]

In general, $e_i(t)$ and $e_o(t)$ are not sinusoidal. Let $E_i(S)$ and $E_o(S)$ represent their Laplace transforms. Equations 4 and 6 become

$$E_o(S) = G(S)E_i(S) \qquad (10)$$

$$G(S) = H(S + j\omega_c) \qquad (11)$$

Transient Performance of Modulating and Demodulating Devices

MODULATORS

An ideal modulating device converts instantly and proportionately an input quantity to the envelope of a sinusoidal carrier wave, which bears a constant phase difference from a given reference. Since the output quantity is the complex envelope, its transfer function is simply a complex constant K. Some of the actual modulators are studied in the following:

Balanced Vibrator and Potentiometer Circuit

While these types differ from an ideal modulator in wave form and in continuity, they can be considered ideal as far as the transient performance is concerned.

Synchro Transmitter and Control Transformer

The synchro transmitter impresses on the control transformer an alternating magnetic field aligned to its own rotor position. While there may be a small transient error between the rotor position of the synchro transmitter and the magnetic field position of the control transformer, such an error does not enter into the closed-loop characteristic and can be neglected as far as stability is concerned.

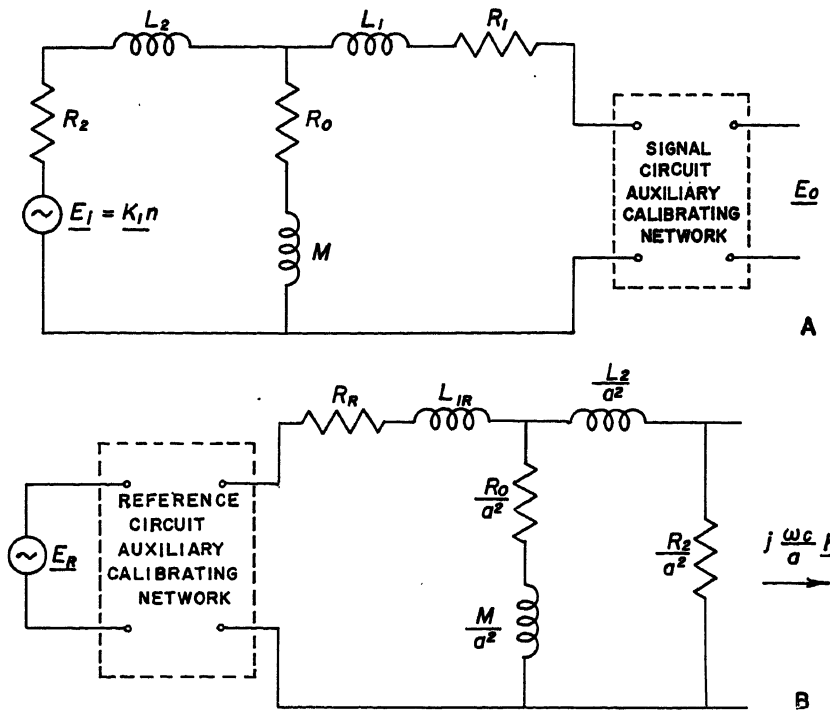If the rotor position of the control transformer is not perpendicular to its magnetic field, the envelope λ of the total

Fig. 1. Transient equivalent circuit of an a-c tachometer

The reference winding impedances together with the auxiliary calibrating network connected to it affect the tachometer performance by modifying K_1, Fig. 1(B). However, these elements do not influence the transient performance directly.

With the complex transform method described in the foregoing section, the equivalent circuit of Fig. 1(A) is analyzed to obtain the transient performance of tachometers with a few typical terminations. The calibrating networks described in Frazier's paper do not introduce any time constant larger than $1/\omega_c$. However, considerable time delay would be introduced by a terminating condenser tuned in resonance with the total signal winding inductance $L_1 + M$, to boost the output signal.

DEMODULATORS

The idealized demodulator characteristic is to generate an instantaneous quantity proportional to the envelope of one component of the input signal while being totally unaffected by the out-of-phase component. Mathematically, regarding the over-all transfer function, its representative operation is to multiply by a complex constant $\bar{K}'$, and to take the real component. There are two types of commonly used demodulating devices.

Phase-Sensitive Detector

The response of a phase-sensitive detector can be expressed as a power series in the reciprocal of the magnitude of the reference voltage. The zero'th order term is proportional to the envelope of the in-phase component of the input signal. The envelope of the out-of-phase component and the time-derivative terms appear in the first order or higher. Thus a phase-sensitive detector can be
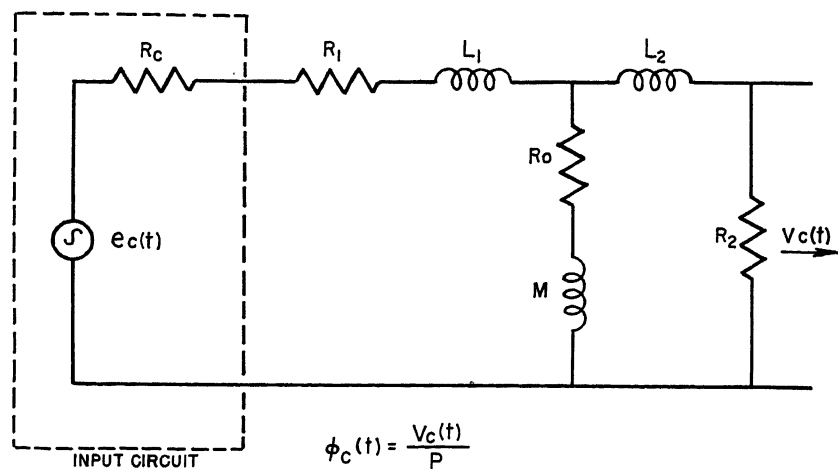
flux linkage to the rotor winding is proportional to the error angle ϵ

$$\lambda(t) = K\epsilon(t) \tag{12}$$

The envelope $e(t)$ of the induced voltage is

$$e(t) = (p_e + j\omega_c)\lambda(t) = K(p_e + j\omega_c)\epsilon(t) \tag{13}$$

Hence, the data transfer function of the synchro control transformer is

$$\frac{E(S)}{\epsilon(S)} = K(S + j\omega_c) = \frac{K}{j\omega_c}\left(1 - j\frac{S}{\omega_c}\right) \tag{14}$$

A-C Tachometer

The steady-state performance of a-c tachometers has been thoroughly analyzed by Frazier.[4] However, with respect to its transient performance, Frazier's published work is limited to determining the characteristic equation of the instantaneous value of the output voltage for the bare tachometer without considering the effects of linearity correction networks. In the Appendix solution of the transient equations is derived with the terminating networks taken into consideration. While the exact solution is too tedious for practical application, a good and simple approximation can be obtained if either of the following conditions hold:

1. Rotor resistance in signal winding terms is large compared to a certain impedance ($a^2 Z_r$ and Z_c, Fig. 7).

2. Speed is small compared to synchronous speed.

These conditions are the same as the conditions for good linearity and are expected to hold very well in practical cases. The transient performance of the tachometer can be represented as an ideal modulator terminated in a linear network as shown in Fig. 1(A). (In Fig. 1 vectors are indicated by a bar under the symbols.) The complex gain constant K is equal to the phasor representing the magnitude and phase of the net alternating flux linking the rotor winding along the reference phase in webers. The terminating network consists of the tachometer equivalent circuit of the signal winding together with the auxiliary calibrating network, if any, in the signal winding circuit.

Fig. 2. Equivalent input network of a 2-phase servo motor

Chang—Transient Analysis of A-C Servomechanisms

Fig. 3. Typical a-c servo and its equivalent d-c servo

$$G_a(S) = \frac{Q_o(S)}{Q_i(S)}$$

made as close to an ideal demodulator as desired.

Two-Phase Servomotor

The response of a 2-phase servomotor is not linear if the motor speed is a substantial fraction of its synchronous speed. However, it is approximately linear for small speeds. The linearity as well as its stability improves with increased rotor resistance.

Generally, the frequency at which the open-loop gain is zero decibel is considerably higher than the highest data frequency so that the error can be small. The electrical transients have appreciable effect only in this high-frequency region. To study the stability in the high-frequency region, it is adequate to assume a small signal superimposed on a slowly varying large signal.

For the case of low speed, as well as for the case of a superimposed small signal, the motor response is linear. It is shown in the Appendix that an actual motor can be considered as a Nichol's simplified motor with a terminating network as shown in Fig. 2. The electrical transients depends entirely on the signal winding circuit and are not appreciably affected by the reference winding circuit.

The transfer function of the 2-phase servomotor can be written into two factors. The first factor gives the ratio of the signal phase net flux linking the rotor to the grid input of the power amplifier. It is a function of the total time derivative operator P and can be integrated into the transfer function of the preceding electric networks. The second factor converts the complex envelope of the signal phase flux to torque. It is to multiply the complex envelope by $K'[1-j(s/(2\omega_c))]$ and take the real component.

Analysis of a System by Complex Transform

The signal in an a-c servo is transmitted directly in part of the loop, as shown by the solid line in Fig. 3(A), and is transmitted in the form of a carrier wave in the other part of the loop, as shown by the dotted line. For a typical position control system, q_i represents the error angle and q_o the motor torque if the demodulator is a 2-phase motor, or q_o represents a voltage if the demodulator is a phase-sensitive detector.

Let $Q_o(S)$ and $Q_i(S)$ be the Laplace transforms of q_o and q_i respectively. If it is possible to find the ratio

$$G_a(S) = \frac{Q_o(S)}{Q_i(S)} \tag{15}$$

then the problem is reduced to a d-c servo problem, as shown in Fig. 3(B), and can be considered solved.

To determine $G_a(S)$, ideal modulator and demodulators are assumed at the moment. The output from the modulator is a voltage $e(t)$

$$e(t) = Kq_i(t) \sin(\omega_c t + \varphi) \tag{16}$$

where K and φ are constants. The complex Laplace transform of this function can be written as

$$E(S) = K e^{j\varphi} Q_i(S) \tag{17}$$

If the demodulator is sensitive to the sine component of the carrier wave only, the Laplace transform of the output q_o of the demodulator is

$$Q_o(S) = \text{real component of } [KK' e^{j\varphi} H(S + j\omega_c) Q_i(S)] \tag{18}$$

Equation 18 follows from the preceding sections. Therefore

$$G_a(S) = \frac{Q_o(S)}{Q_i(S)} = KK' \times \text{real component of}$$
$$[e^{j\varphi} H(S + j\omega_c)] \tag{19}$$

In general the modulator and demodulator have energy storage elements, and equation 19 can be modified as

$$G_a(S) = K_1 K_2 \times \text{real component of}$$
$$[H(S + j\omega_c) F(S) e^{j\varphi}] \tag{20}$$

where $H(S) = H_1(S) \quad H_2(S) \quad H_3(S)$. $H_2(S)$ is the transfer function of the explicit network. $H_1(S)$ depends on the modulator only. $H_3(S)$ and $F(S)$ depend on the demodulator only. They are:

Balanced vibrator or potentiometer: $H_1(s) = 1$

Synchro transmitter—control transformer: $H_1(S) = S$

A-c tachometer: $H_1(S) = $ transfer function of network of Fig. 1(A).

Phase-sensitive detector: $H_3(S) = 1 \; F(S) = 1$

Two-phase servomotor: $F(S) = 1 - j[S/(2\omega_c)]$

$H_3(S) = \frac{1}{S} \times$ transfer function of network of

Fig. 2

It follows from these considerations that the transient effects of various a-c servo components cannot be treated separately except as complex mathematical entities. They cannot be determined experimentally as an independent effect. A simple analogy to this situation is that the real component of a product of complex factors does not equal to the product of the real components of the constituent factors.

Mathematical Approximations

The function $H(S + j\omega_c)$ can be separated into a steady-state factor $H(j\omega_c)$ and a transient factor depending on S. Generally $H(S)$ is composed of two types of factors, denoted by $H_a(S)$ and $H_b(S)$

$$H_a(S) = 1 + TS \tag{21}$$

$$H_b(S) = 1 + \frac{2a}{\omega_1} S + \frac{S^2}{\omega_1^2}, \text{ where } a < 1 \tag{22}$$

Physically, an incidental time-dependent circuit, e.g., a phase-shift circuit, or a motor equivalent circuit, gives rise to factors of the form $H_a(S)$ while a circuit intentionally inserted to stabilize the servo gives rise to factors of both the forms $H_a(S)$ and $H_e(S)$. Thus

$$H_a(S + j\omega_c) = 1 + T(S + j\omega_c) = H_a(j\omega_c)$$
$$(1 + T_1 S) \tag{23}$$

MAXIMUM SINUSOIDAL RESPONSE Mm

(graph values and labels)

9.8 8.4 $W_1=0.9WC$ $W_1=1.1WC$ 10 8.7 $W_1=0.9WC$ $W_1=1.1WC$ 10.2 9.2 10.5 9.9 $W_1=WC$ 9.7 10.4 11.2 12.1 13.7

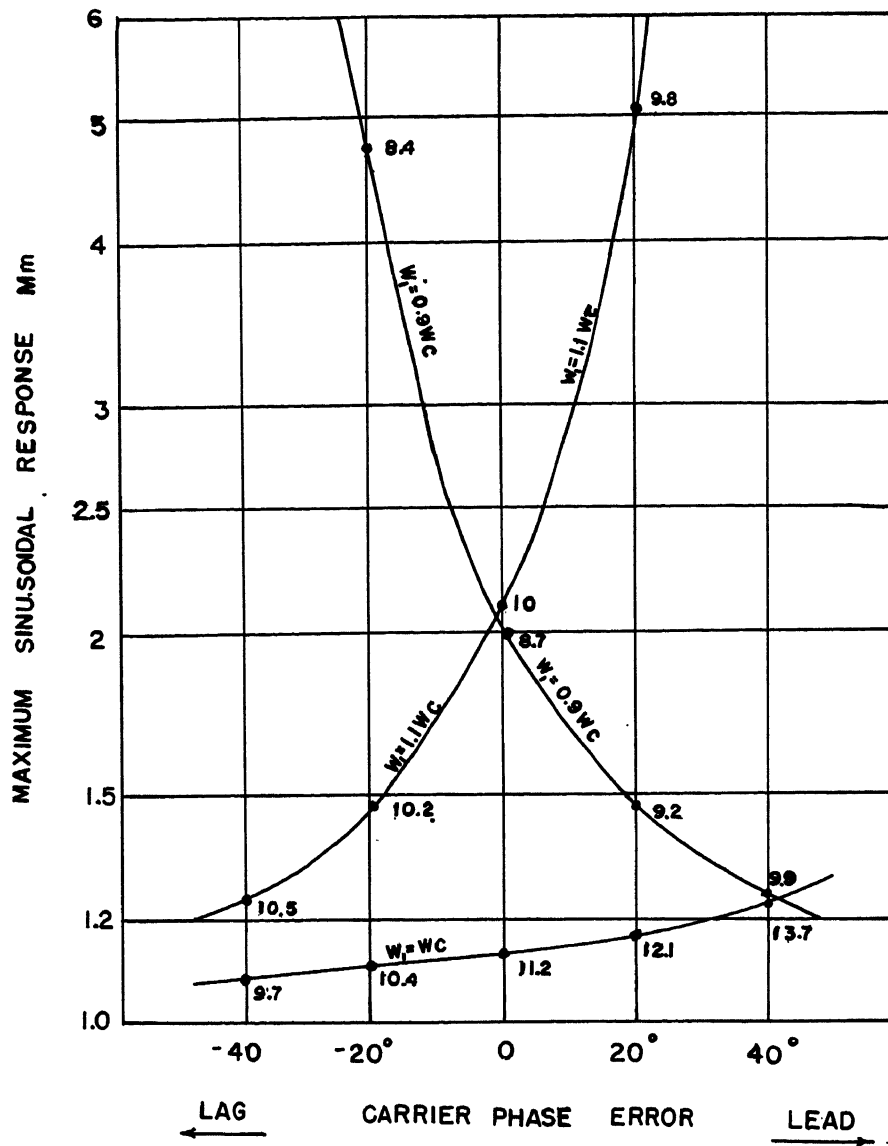-40 -20° 0 20° 40°

LAG ← CARRIER PHASE ERROR → LEAD

Fig. 4. Maximum modulus of typical 60-cycle a-c servomechanism with detuned lead network and carrier phase error

$H_b(S+j\omega_c)=H_b(j\omega_c)(1+T_aS)(1+T_2S)$ (24)

where

$$T_1=\frac{T}{1+j\omega_c T}$$ (25)

$$T_a=\frac{1}{a\omega_1+j(\omega_c-\omega_1\sqrt{1-a^2})}$$ (26)

$$T_2=\frac{1}{a\omega_1+j(\omega_c+\omega_1\sqrt{1-a^2})}$$ (27)

The absolute values of T_1 and T_2 are always smaller than $1/\omega_c$. T_a can be many times larger than $1/\omega_c$ if a is small and $\omega_1=\omega_c$.

Generally, both types of time constants are present in $H(S+j\omega_c)$. A useful approximation is to separate the factors into two groups. The first group consists of factors with time constants T_a and T_b larger than $1/\omega_c$; the second group consists of factors with time

constants. T_1, T_2, ... smaller than $1/\omega_c$. The second group of factors is expanded into a power series. For example

$$\frac{(1+T_aS)(1+T_bS)(1+T_1S)}{(1+T_2S)(1+T_3S)(1+T_4S)}$$
$$=(1+T_aS)(1+T_bS)[1+(T_1-T_2-T_3-T_4)S+...]$$ (28)

The foregoing mathematical manipulations have two points of practical significance, namely:

1. The separation of $H(S+j\omega_c)$ into a steady-state factor and an electrical transient factor allows the steady-state factor $H(j\omega_c)$ to be determined experimentally with the transient factor used as a modification. This procedure is very convenient especially in describing the rotating components.

2. It is known that, with a sinusoidal data input, the effects of nonresonant networks are essentially a phase shift in the carrier factor and a phase shift in the

envelope factor. The carrier phase shift is taken into account by the steady-state factor $H(j\omega_c)$. The data time constants T_1, T_2, ... for a nonresonant network are always less than $1/\omega_c$.

Equation 28 takes into account the combined envelope phase-shift effect.

An Example of Analysis

A typical 60-cycle position control servomechanism with frequency and phase errors is analyzed to illustrate the method. It has the following elements:

1. Synchro control transformer.
2. Amplifier, with a 2-stage resistance-capacitance coupling network, a phase-shift network, and a carrier frequency lead network.
3. Two-phase servomotor.

Their respective transfer functions are:

$$H_1(S)=S$$

$$H_{2a}(S)=\left(\frac{0.03S}{1+0.04S}\right)^2$$

$$H_{2b}(S)=\frac{1}{S}$$

$$H_{2c}(S)=\frac{S^2+\frac{2}{15}\omega_1S+\omega_1^2}{S^2+4\omega_1S+\omega_1^2}$$

$$H_3(S)=\frac{0.00152}{(1+0.00447S)(1+0.000435S)}$$

$$G_d=\frac{K_3}{S(1+0.2S)}$$

The transfer function $G_a(S)$ can be determined from equation 20 with the help of the approximate equation 28 for various carrier frequency conditions. They are

For $\omega_1=\omega_c$

$G_a(S)=K\times$ real component of

$$\left[e^{j\varphi}\left(1+15\frac{S}{\omega_c}\right)\times\left(1+1.44\underline{/148.1°}\frac{S}{\omega_c}\right)\right]$$ (29)

For $\omega_1=1.1\omega_c$

$G_a(S)=K\times$ real component of

$$\left\{1.82e^{j(\varphi-53°)}\left[1+8.1\underline{/64.0°}\times\left(\frac{S}{\omega_c}\right)-11.6\underline{/21.8°}\left(\frac{S}{\omega_c}\right)^2\right]\right\}$$ (30)

There is a similar equation for $\omega_1=0.9\omega_c$.

Fig. 4 illustrates the calculated maximum modulus to a sinusoidal data input at various values of actual carrier phase error. It is assumed that the gain of the amplifier is adjusted so that the over-all gain coefficient of the system is always 600 radians per second. Fig. 5 gives

the transient response to a unit step input under the condition that $\omega_1 = 1.1\omega_c$ and a carrier phase error of -20 degrees, ($\phi = 33$ degrees). Fig. 6 illustrates the situation if the carrier frequency is drifted away from its initial value without further adjustment made on the system. The over-all gain and time constants are plotted as a function of the initial phase-angle error ϕ.

In Fig. 5 the response function is not plotted for small values of t for the following reason: Since the complex transform method is limited to data frequencies lower than the carrier frequency, the calculation for a unit step input is not strictly correct. For small values of t, the high-frequency terms are predominant, and the result is erroneous. For larger values of t, the high-frequency terms are subdued by destructive interference and the calculated result is expected to hold. Fig. 5 is a plot of equation 57 for values of t larger than 0.0167 second, corresponding to one period of the carrier. Physically, the transient response of the servo is not expected to be consistent within one period of the carrier immediately after the step change is applied.

Appendix. Derivation of Transient Equivalent Circuits of A-C Tachometers and Servomotors

The derivation of the transient equivalent circuits is presented separately in the Appendix for two reasons:

1. It is not necessary to know how the circuits are derived to be able to apply them.
2. The derivation requires knowledge in a-c machinery not related to the rest of the feedback control field.

The basic cross-field equations for the a-c tachometers and servomotors are essentially the same, namely

$$v_{1c} - (R_1 + L_1 p)i_{1c} = Mp(i_{1c} - i_c) \tag{31}$$

$$\phi_c = M(i_{1c} - i_c) - L_2 i_c \tag{32}$$

$$v_{1r} - (R_{1r} + L_{1r}p)i_{1r} = \frac{M}{a^2}p(i_{1r} - i_r) \tag{33}$$

$$\phi_r = \frac{M}{a^2}(i_{1r} - i_r) - \frac{L_2}{a^2}i_r \tag{34}$$

$$R_2 i_c = p\phi_c - na\phi_r \tag{35}$$

$$\frac{R_2}{a^2}i_r = p\phi_r + n\frac{\phi_c}{a} \tag{36}$$

$$\text{Torque} = K\left(a\phi_r i_c - \phi_c \frac{i_r}{a}\right) \tag{37}$$

where
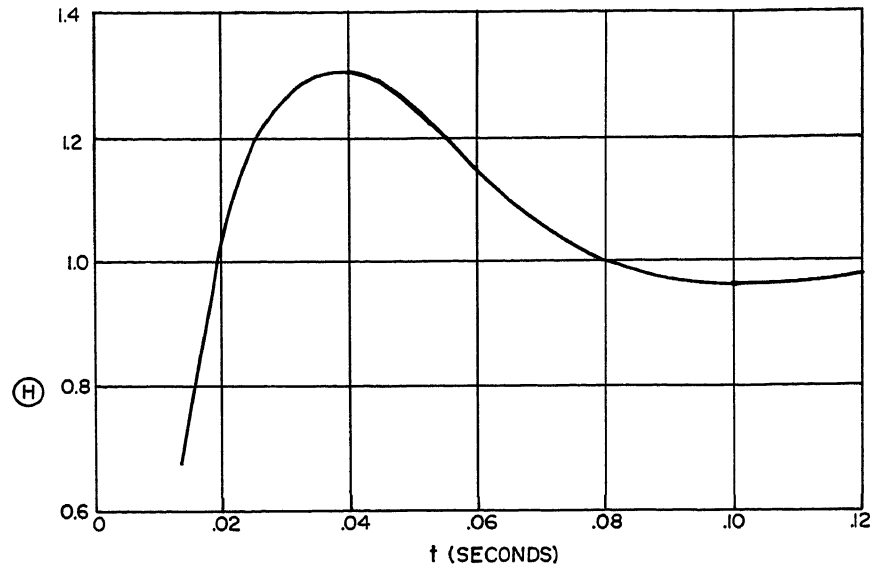
Fig. 5. Transient response to unit step input of typical 60-cycle a-c servomechanism with detuned lead network and 20-degree lagging carrier phase angle
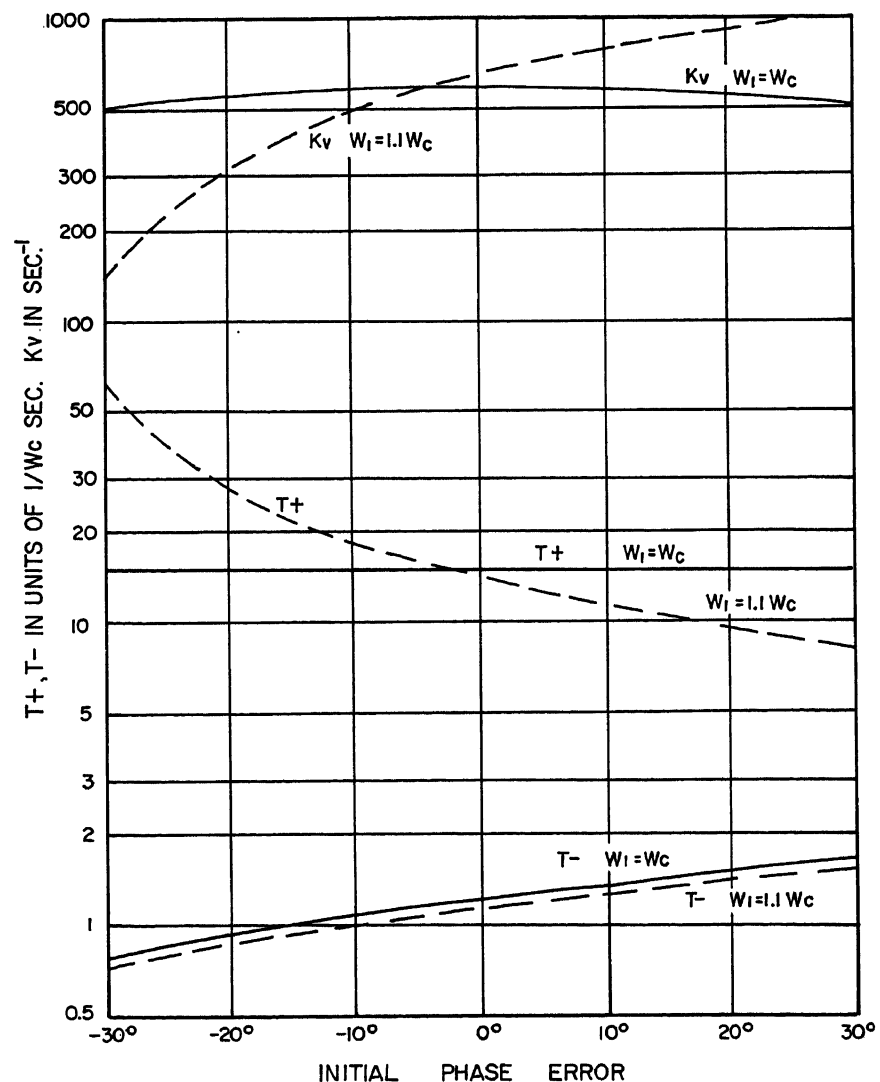
Fig. 6. Time constants and gain as functions of frequency and phase errors

$$G_a(S)G_d(S) = \frac{K_v(1+T+S)}{S(1+T-S)(1+T_mS)}$$
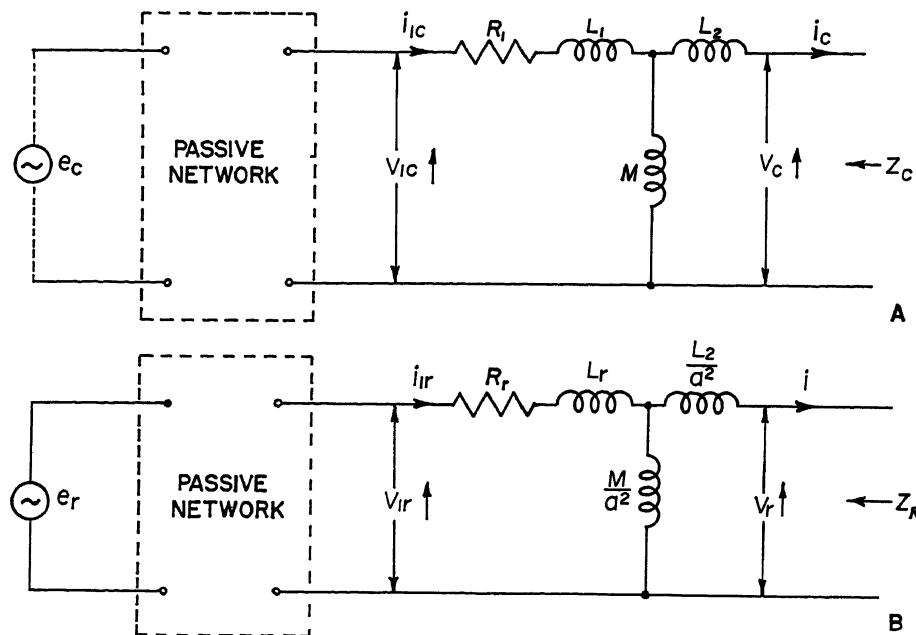
Fig. 7. Equivalent circuits representing basic equations for a-c tachometers and servomotors (cross-field theory)

v_{1c}, i_{1c}, i_c, ϕ_c = control winding terminal voltage, winding current, control winding axis rotor current, and net flux linking the rotor conductors along the control winding axis, all in control winding terms

v_{1r}, i_{1r}, i_r, ϕ_r = the respective reference winding quantities, all in reference winding terms

R_1, L_1, M, L_2, R_2 = control winding resistance, leakage inductance, mutual inductance, rotor leakage inductance, and rotor resistance, all in control winding terms

R_{1r}, L_{1r} = reference winding resistance and leakage inductance

a = ratio of effective control winding turns to effective reference winding turns

For the a-c tachometers, the word "control" should be replaced by the word "signal," and the choice of the reference direction of i_{1c} and i_c in this paper is opposed to that of the literature in general.

Equations 31 and 33 give the effects of winding resistances, leakage, and mutual inductances. Equations 32 and 34 give the net rotor fluxes in terms of currents. Equations 35 and 36 give the rotor currents in terms of transformer and speed electromotive forces. Equation 37 gives the motor torque as interaction between fluxes and currents. Defining v_r and v_c as $d\Phi_r/d_t$ and $d\Phi_c/d_t$ respectively and differentiating equations 32 and 34 with respect to t

$$v_c = Mp(i_{1c} - i_c) - L_2 p i_c \quad (38)$$

$$v_r = \frac{Mp}{a^2}(i_{1r} - i_r) - \frac{L_2 p}{a^2} i_r \quad (39)$$

Equations 31, 33, 38, and 39 are represented by Fig. 7.

As resonant effects are not present, the electrical transients introduce data time constants of the order of $1/\omega_c$. These effects are appreciable only in the relatively high frequency range, where the data

frequency is a substantial fraction of ω_c. Hence the electrical transients are important in two respects:

1. In determining the servo response to high-frequency components of the reference input.
2. In determining the servo stability.

Generally, the power spectrum of the reference input is concentrated in a frequency range considerably lower than the frequency of zero decibel open-loop gain. This is a necessary condition for low error. Consequently the high-frequency components in the reference input are of relatively small magnitude. For both of the purposes just listed, the rapid data inputs can be considered as a small variation imposed on a possibly large slow data input for which the fluxes and currents in the servomotor or tachometer are already in steady-state balance. Towards the superposed small rapid variations the response of the motor and tachometer is approximately linear and can be treated in the manner now described.

A-C Tachometers

For a-c tachometers, e_c in Fig. 7(A) represents the measured instantaneous output voltage across the terminals of the calibration network, and the flow of energy is from right to left. Without changing the direction of i_c in the representation, e_c and v_c can be expressed as

$$e_c(t) = -Z_T(p)i_c(t) \quad (40)$$

$$v_c(t) = -Z_c(p)i_c(t) \quad (41)$$

The impedances Z_T and Z_c can be calculated from Fig. 7(A). Similarly, from the circuit of Fig. 7(B), the current is expressed as

$$i_r(t) = Y_u(p)e_r(t) - \frac{1}{Z_r(p)} v_r(t) \quad (42)$$

where $Y_u(p)$ is the transfer admittance i_r/e_r with the right-hand terminal of Fig. 7(B) short-circuited. $Z_r(p)$ is the terminal impedance with e_r replaced by a short circuit. All impedances in equations 40, 41, and 42 are in operator forms.

For a small change in speed δn, the corresponding changes in fluxes and currents are obtained by taking differentials of equations 35, 36, 41, and 42. They are

$$R_2(\delta i_c) = p(\delta\phi_c) - (\delta n)a\phi_r - na(\delta\phi_r) \quad (43)$$

$$\frac{R_2}{a^2}(\delta i_r) = p(\delta\phi_r) + (\delta n)\frac{\phi_c}{a} + n\frac{(\delta\phi_c)}{a} \quad (44)$$

$$p(\delta\phi_c) = -Z_c(p)(\delta i_c) \quad (45)$$

$$p(\delta\phi_r) = -Z_r(p)(\delta i_r) \quad (46)$$

Substituting equations 45 and 46 in equations 43 and 44

$$[R_2 + Z_c(p)](\delta i_c) - \frac{na}{p} Z_r(p)(\delta i_r) = -a\phi_r(\delta n) \quad (47)$$

$$\frac{an}{p} Z_c(p)(\delta i_c) + [R_2 + a^2 Z_r(p)](\delta i_r) = a\phi_c(\delta n) \quad (48)$$

Solving (δi_c) from equations 47 and 48 gives

$$(\delta i_c) = \frac{-a\phi_r[R_2 + a^2 Z_r(p)] + \phi_c \dfrac{na^2}{p} Z_r(p)}{[R_2 + Z_c(p)][R_2 + a^2 Z_r(p)] + \dfrac{n^2 a^2}{p^2} Z_r(p) Z_c(p)} (\delta)n \quad (49)$$

In equation 49, ϕ_r and ϕ_c are the steady-state values of reference axis flux and control axis flux respectively.

Under steady-state conditions, equations 35, 36, 40, 41, and 42 can be solved for I_c, ϕ_c, and E_c. They are

$$I_c = -\frac{a\phi_r}{R_2 + Z_c} n \quad (50)$$

$$\phi_c = \frac{V_c}{j\omega_c} = \frac{Z_c}{R_2 + Z_c} \frac{n}{j\omega_c} (a\phi_r) \quad (51)$$

$$E_c = \frac{Z_T Y_u R_2 Z_r a E_r}{\left[(R_2 + Z_c)(R_2 + a^2 Z_r) - a^2 Z_r Z_c \dfrac{n^2}{\omega_c^2}\right]} \frac{n}{j\omega_c} \quad (52)$$

Equation 52 is a generalized version of Frazier's solution to the steady-state problem. Frazier's methods of linearizing the performance of a-c tachometers are simply means of reducing Z_r and Z_c. Equations 50, 51, and 52 are vector equations with the capitalized quantities being the corresponding vectors of the instantaneous quantities. In the impedances and admittances, the operator p is replaced by $j\omega_c$.

The steady-state solutions provide a yardstick for simplifying equation 49. From equation 51, it is evident that the second terms of both the denominator and numerator of 49 are of the order

$$\frac{a^2 Z_r Z_c}{(R_2 + a^2 Z_r)(R_2 + Z_c)} \frac{n^2}{\omega_c^2}$$

compared to the respective first terms.

This is exactly the factor which has to be small for linear steady-state response. In practically satisfactory tachometers, the second terms can be neglected. Equation 49 becomes

$$(\delta i_c) = \frac{a\phi_r}{R_2 + Z_c(p)} (\delta n) \qquad (53)$$

Equations 53 and 40 are represented by the equivalent circuit of Fig. 1. The constant K_1 is simply $a\phi_r$. However, it is not necessary to calculate ϕ_r. The equivalent circuit can be used to determine the electrical transient factor and attach it to the constant calibrated in steady-state condition as shown in the section entitled "An Example of Analysis." The constant R_o is a small correction factor for iron loss.

Servomotors

Equations 40 and 41 are replaced by

$$i_c(t) = Y_v(p)e_c(t) - \frac{1}{Z_c(p)} v_c(t) \qquad (54)$$

For small rapid variations in the control phase electromotive force e_c, the variations in speed are small and produce an effect of still smaller order. To obtain correlations between the variations in currents and fluxes, the speed is treated as a constant. Another assumption is that, as the reference winding is connected directly to a source of electromotive force e_r, the reference phase flux ϕ_r is determined almost entirely by e_r and its variation resulting from a small variation in e_c can be neglected. Taking the variations from equations 35, 36, and 37 one obtains

$$R_2(\delta i_c) = p(\delta \phi_c) = (\delta v_c) \qquad (55)$$

$$\frac{R_2}{a^2}(\delta i_r) = n\frac{(\delta \phi_c)}{a} \qquad (56)$$

$$\delta \text{ Torque} = K\left[a\phi_r(\delta i_c) - (\delta \phi_c)\frac{i_r}{a} - \phi_c\frac{(\delta i_r)}{a} \right]$$
$$= \frac{K}{R_2}[a\phi_r p(\delta \phi_c) - (\delta \phi_c)ap\phi_r - n\delta(\phi_c^2)] \qquad (57)$$

Equations 54 and 55 are represented by the equivalent circuit of Fig. 2. Equation 57 gives rise to the factor $1 - j.5(S/\omega_c)$ as shown in the following

$$\delta\phi_c = \delta\phi_1(t) \sin \omega_c t + \delta\phi_2(t) \cos \omega_c t \qquad (58)$$

$$p(\delta\phi_c) = [p_e(\delta\phi_1) - \omega_c(\delta\phi_2)] \sin \omega_c t + [p_e(\delta\phi_2) + \omega_c\delta\phi_1] \cos \omega_c t \qquad (59)$$

$$\phi_r = \phi_r \cos \omega_c t, \ p\phi_r = -\omega_c\phi_r \sin \omega_c t \qquad (60)$$

Substituting equations 58, 59, and 60 in 57 and neglecting the double frequency terms, one obtains

$$\delta \text{ Torque} = \frac{K}{R_2}\left\{ a\phi_r\left[\omega_c(\delta\phi_1) + \frac{1}{2} p_e(\delta\phi_2) \right] - n[\phi_1(\delta\phi_1) + \phi_2(\delta\phi_2)] \right\} \qquad (61)$$

The term $n(\phi_1\delta\phi_1 + \phi_2\delta\phi_2)$ is due to internal damping of the control winding flux. It is small compared to $\omega_c a\phi_r(\delta\phi_1)$. Retaining only the large terms, one obtains

$$\text{Torque} = \frac{K}{R_2}\left\{ a\phi_r\left[\omega_c(\delta\phi_1) + \frac{1}{2} p_e(\delta\phi_2) \right] \right\} \qquad (62)$$

Equation 62 can be expressed in complex transform notation as

$$\text{Torque} = \frac{Ka\phi_r\omega_c}{R_2} \text{ real} \left[\left(1 - \frac{j}{2}\frac{p_e}{\omega_c} \right) \times (\delta\phi_1 + j\delta\phi_2) \right] \qquad (63)$$

The factor $[1 - (j/2)(p_e/\omega_c)]$ is equivalent to an expression first derived by Nichols, (reference 1).

References

1. STABILIZATION OF CARRIER FREQUENCY SERVOMECHANISMS, A. Sobczyk. *Journal*, Franklin Institute, Philadelphia, Pa., July, Aug., Sept. 1948.

2. EFFECTS OF CARRIER SHIFTS ON DERIVATIVE NETWORKS FOR A-C SERVOMECHANISMS, George M. Attura. *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 612–18.

3. TRANSFER FUNCTION FOR A 2-PHASE INDUCTION SERVO MOTOR, Lloyd O. Brown, Jr. *Ibid.*, pt. II, pp. 1890–93.

4. ANALYSIS OF THE DRAG-CUP A-C TACHOMETER BY MEANS OF 2-PHASE SYMMETRICAL COMPONENTS R. H. Frazier. *Ibid.*, pp. 1894–1906.

5. THEORY OF SERVOMECHANISMS (book), H. M. James, N. B. Nichols, R. S. Phillips. McGraw-Hill Book Company, Inc., New York, N. Y., 1947.

6. PRINCIPLES OF SERVOMECHANISMS (book), G. S. Brown, D. P. Campbell. John Wiley and Sons, Inc., New York, N. Y., 1948.

7. TRANSIENTS IN LINEAR SYSTEMS (book), M. F. Gardner, J. L. Barnes. *Ibid.*, 1942.

8. NETWORK SYNTHESIS BY GRAPHICAL METHODS FOR A-C SERVOMECHANISMS, George A. Bjornson. *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 619–25.

9. OPERATING CHARACTERISTICS OF 2-PHASE SERVOMOTORS, R. J. K. Koopman. *Ibid.*, vol. 68, pt. I, 1949, pp. 319–28.

Thermal Stability of a New Insulating Material Used in Traction Motors

R. W. FINHOLT
NONMEMBER AIEE

Properties Needed in Traction Motor Insulation

MODERN traction motors operate under severe conditions. The latest adopted standard for traction motors lists them as class B.[1] This provides that the peak value by resistance of the armature shall not exceed 160 degrees centigrade (C), and the peak value for the fields shall not exceed 170 C. Modern traction motors on heavy Diesel-electric locomotives often exceed these temperature limits by 20 to 30 C. The temperatures in the hottest part of the armature and the hottest part of the field will exceed these average resistance measurements.

Coupled with these high temperatures are severe vibration conditions, dust, water, varying humidity, highly variable maintenance, and shock. Most traction-motor equipment operates on 3,000 volts or less on direct current, or 1,500 volts or less on alternating current. The railroads expect and get long motor life coupled with a very high degree of dependability.

A general problem in insulation is that of high temperature combined with a number of mechanical destructive factors. In the search for better materials, tests have been run on the high-temperature materials that have become available in the last 15 years. One of these in particular has proved to have such

excellent properties, both physically and economically, that the General Electric Company is adopting it widely for use on traction motors. This new material is known as mica mat.

Mica-Mat Insulation

GENERAL PROPERTIES

The form of mica mat found most useful in traction-motor insulation is impregnated with silicone resin. This is available as a 4.5-mil sheet material, and also as a 5-mil glass-backed tape in widths up to 38 inches. Other types of impregnants were rejected after heat-aging tests described in this paper. The material finally adopted was the silicone-treated General Electric mica mat. It is used both as wrappers on armature coils and as tape on field coils. After

Table I. Electrical Properties of Mica Mat

	Temperature, C	Untreated		Silicone Treated	
		0.002 Inch Thick	0.004 Inch Thick	0.002 Inch Thick Plus Glass Cloth (0.0018 Inch Thick)	0.004 Inch Thick Plus Glass Cloth (0.0018 Inch Thick)
Dielectric constant, 60 cycles	{ 25	1.6	1.6		2.3
	100			2.2	2.4
	125				2.3
Power factor	{ 25	0.08%	0.08%	2.3%	2.5%
	100	0.21%	0.23%	1.7%	2.5%
	125			1.8%	2.7%
Dielectric strength (short time)	25	700 volts per minute*	700 volts per minute*	450 volts per minute†	450 volts per minute†

* 2-inch electrode. † 1/4-inch electrode.

Table II. Physical Properties of Mica Mat*

Properties	Measurements	
Thickness, inches	0.004	reference 2
Weight per square foot, grams	17	
Tensile strength minimum, pounds per 1 inch	10	reference 3
Folding endurance, flexes	10	Massachusetts Institute of Technology folding endurance tester (load 0.2 kilograms)
Compressive strength at 200 C, minimum, psi	10,000	
Binder content, % by weight	11 to 17	
Carbon content, %	5.8 to 9.0	reference 4

* General Electric Specification *A14F7D1*. General Electric Chemical Department material *77862*, silicone-treated mica-mat sheet.

Table III. Effect of Heat Aging on Dielectric Strength of Silicone-Treated Mica Mat

3¹/₄ Wraps of Mica Mat (13¹/₂% Silicone) on Copper Bars 6 Inches by 0.200 Inch by 0.263 Inch

Sample No.	Time at 250 C	Average Dielectric Strength, Volts (4-Inch Section, Short Time)
1	0	7,875
2	1 week	8,250
3	2 weeks	7,950
4	1 month	7,925
5	2 months	8,520
6	4 months	7,550
7	14 months	7,670

working with various types of resins different types of mica mat, different kinds of backing material, and varying types of applications, the forms of mica mat used were narrowed down to the two just mentioned. Their general properties are listed in Tables I and II. Typical breakdown values determined by tests on varying thickness of material applied to bars in a manner similar to that used in service are shown in Fig. 1.

GENERAL TESTS

An insulation material is usually subjected to four types of tests before
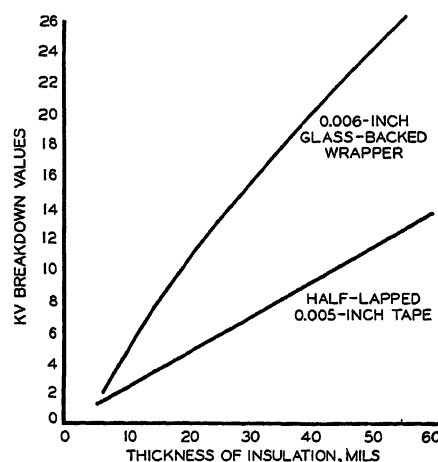
Fig. 1. Dielectric strength of silicone-treated mica tape and sheet on copper conductor

it is accepted as a standard manufacturing material.

1. Laboratory evaluation tests to determine the fundamental properties of the material.

2. Service-type tests on actual equipment or dummy mockups.

3. Field tests with full-scale equipment.

4. Acceptance and factory standards tests needed for actual manufacture if the first three tests are satisfactory.

With mica mat the preliminary laboratory tests included humidity and pressure tests, heat aging, various electrical tests, tensile strength, flexural strength, bursting strength flexing over a sharp-edged mandrel, resin content, thickness, density, and exposure to arcing and corona. The second type of testing demonstrated that mica mat insulated motors could be operated successfully under severe conditions. Motors were then made in trial lots for the third type of testing and put out in selected traction service under field conditions. After a year of field experience it became evident that mica mat was a remarkably good insulating material. Accordingly, factory production was started on several lines of motors.

During the time that field tests were in progress, co-operative work was carried out on the manufacture of mica mat to improve its ultimate physical properties. The fourth type of testing

was then undertaken to set up acceptance tests and manufacturing standards. Approximately 1 year after the setting up of satisfactory tests and standards, mica mat was accepted as a standard insulating material to be used by our design engineers wherever its application appeared proper.

Of all the testing mentioned here, only heat testing will be discussed in detail in this paper; but other tests and heat tests cannot be divorced entirely. If mica mat had proved to be deficient in any one respect, e.g., in the ease of application to armature coils, it could not have been used as a standard manufacturing material.

HEAT TESTS

The first tests were on flat sheet heataged at 250 C for 5 days. Mica-mat sheets impregnated with alkyd resin, with silicone, and with several polyesters were tested. At the end of 1 week only the sheet that had been impregnated with silicone retained its original properties. The other sheets had become brittle, mechanically weak, and had all but lost their original dielectric strength.

A long-term aging test was run in which short lengths of copper wire 0.200 inch by 0.265 inch were wrapped with 3¹/₄ wraps of mica mat. These were put in an air oven at 250 C. As can be seen from Table III, even after 14 months there was no significant change in the short-time dielectric properties of the material. Flat control sheets of unimpregnated mica mat likewise showed no change in dielectric after over 1 year at 250 C. When the dielectric strength of the silicone-treated mica mat was not impaired by short-time heating at 250 C, attempts were made to accelerate heat degradation. The temperature was raised to 300 C, then 350 C, and finally, after no degradation at those tempera-

Fig. 2. Effect of
600 C on silicone-
treated mica mat.
Samples heated for
3½ hours at 600 C
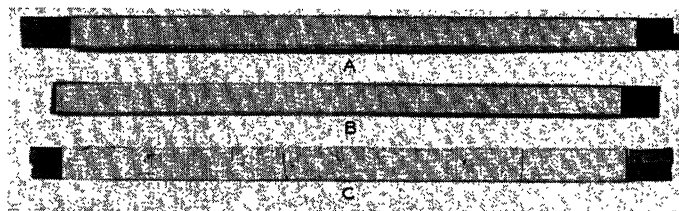
A—Mica mat and silicone, catalyzed.
8,000 volts before heat; 8,000 volts after
heat
B—Mica mat and silicone, uncatalyzed.
8,750 volts before heat; 7,000 volts after
heat
C—Asbestos and silicone, catalyzed.
4,125 volts before heat; >1,000 volts
after heat

tures, several samples were put in a furnace and held at 600 C for 3½ hours. At the end of this time the copper wire was badly oxidized and a comparison sample of silicone-impregnated asbestos was falling apart. As can be seen in Fig. 2, the mica-mat samples were in good condition, were mechanically intact, and still had practically all of their original dielectric strength. The wrappers were cut off and analyzed for silicone by determining the per cent (%) of carbon. This analysis showed only 0.218% of carbon which, after subtracting the 0.14% value of the blank control, corresponded to less than 0.2% of silicone.

The silicone-impregnated mica mat profited from its partnership with silicone at all temperatures. Even though the

Table IV. Effect of 100% Relative Humidity on Mica Mat Containing 10% Silicone

Sample No.	Temperature, C	Hours of Exposure	% Change in Dielectric*
1	22	24	+ 4
2	22	96	+ 3
3	50	24	− 6
4	50	96	− 5
5†	22	24	− 3
6‡	50	24	−14

* Reference 5, short time.
† Untreated.
‡ Controls.

Table V. Effect of 100% Relative Humidity with Various Amounts of Silicone Resin

Tests Run at 22 C After 24 Hours' Exposure at 50 C

Sample No.	% Silicone	% Change in Dielectric*
1	0	−7.5 to −14
2	11.9	−4.2
3	21.0	−2.8
4	36.0	+3.2

* Reference 5, short time.

carbon part of the silicone had burned out at the very high temperatures, enough of a binding action had taken place to render the mica mat still strong, still water resistant, and still possessed of its original dielectric strength. This straight thermal testing showed that mica mat had properties far in excess of those actually needed for the application.

HUMIDITY TESTS

Silicone-treated mica mat has good dielectric strength under severe humidity conditions. Tests at humidities from 42 to 90% showed that both silicone-treated mica mat and untreated mica mat had an increase in dielectric strength. Under more severe conditions of 100% humidity at several different temperatures, untreated mica mat suffered a small loss in dielectric strength, but silicone-treated mica mat showed an increase in dielectric strength as long as no liquid water was present; see Tables IV and V. A loss in strength was found when there was condensation of water within the mica-mat sheet. Not until the sheet contained over 30% of silicone did the dielectric strength remain unimpaired by condensed water. This is the highest amount of silicone that can be gotten into the sheet itself. Higher percentages can apparently be obtained, but this is due to layers of silicone on the outside of the sheet. These tests showed that mica mat was satisfactory as far as humidity was concerned. Comparable tests run with silicone-pasted (12% silicone) mica tape on a glass backing gave results almost exactly like mica mat with about 12% silicone.

Heat-aging the mica mat at 250 degrees before humidity tests gave results comparable with nonaged samples. To evaluate severe heat aging, test bars run at 600 C for 3½ hours were put in a 100% humidity chamber at 50 C. After 24 hours they showed a loss of 6% in dielectric strength. Untreated mica mat is very water-sensitive and when dropped into a glass of water can be broken up to a slurry. The 600 C aged material, however, did not break up in water and retained most of its
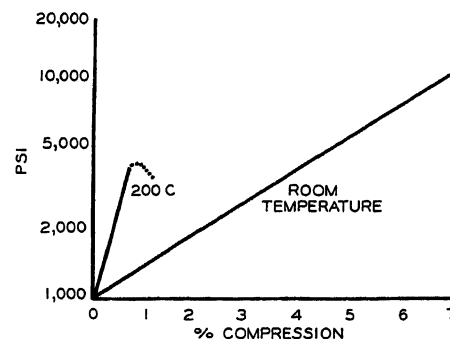
Fig. 3. Comparison of silicone-treated mica mat at various pressures

original nonwetting characteristics. This showed that mica mat impregnated with silicone would retain its water resistance even after prolonged heating.

HEAT AND PRESSURE TESTS

The insulations in the armatures and in the field coils have to be planned to withstand 200 C or more. In service there is also both centrifugal pressure and shock pressure and so the combined effect of heat and pressure on silicone-treated mica mat was evaluated. The apparatus used was a temperature controlled Carver Laboratory press fitted with a hydraulic pressure gauge and two dial gauges to measure jaw movement. The initial height of a stack of 1-inch square samples was measured at a specified pressure with a telescoping gauge and micrometer. It was necessary to calibrate the Carver press because of the significant amount of stretch, play, and compression in the system. Calibration was accomplished with a small steel block in place of the samples. The results in the calibration run at each pressure were subtracted from the reading of the sample run. The samples to be tested were placed as near to the center of the jaws as possible and the pressure raised to 1,000 pounds per square inch (psi). At the end of 5 minutes the dial gauges were zeroed and the initial stack height was measured by taking readings at the four corners of the platen and averaging them. After measuring the initial stack height (approximately ½ inch), the pressure was raised to 1,000 psi every half-minute and readings were taken just before the pressure was raised. Failure points were taken at the first 1,000-pound interval that the sample would not stand.

As shown in Fig. 3, the % compression versus the pressure was a linear logarithmic function at both room temperature and at 200 C. At 200 C the sample treated with 13.5% silicone disintegrated
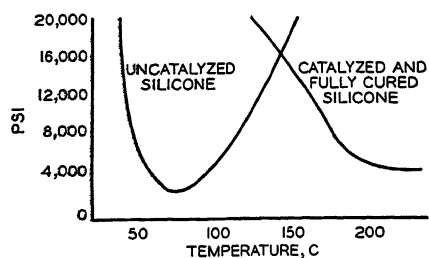
Fig. 4. Pressure causing failure of silicone-treated mica mat at various temperatures
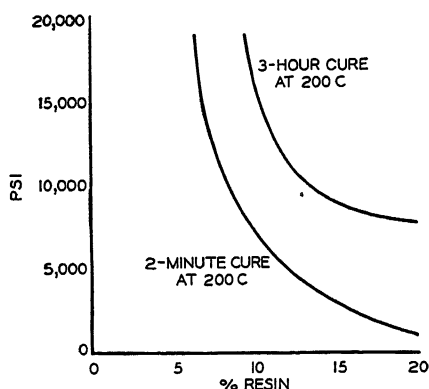

Fig. 5. Pressures causing failure of silicone-treated mica mat at 200 C and changing amounts of resin

at about 4,000 psi. A careful check of pressure-temperature failure points showed that a moderately cured silicone resin at 13.5% resin content had decreasing resistance to pressure with increasing temperatures; see Fig. 4. When an uncatalyzed resin was used as the impregnant, the resin flowed out of the mica mat under heat and pressure. Because of the low content of silicone resin remaining, the crush resistance falsely appeared to be high at high temperatures. An uncatalyzed resin is, of course, extremely undesirable from a crush-resistant standpoint.

Increasing the concentration of silicone resin lowered the crush resistnace. The results of this test are shown in Fig. 5. The lower curve is for a 2-minute cure at 200 C, while the upper curve is for a 3-hour cure at 200 C. These show that below about 7% silicone no failure takes place, even at 20,000 psi; but with increasing amounts of silicone, gradual reduction in the crush resistance takes place.

The general conclusion from the pressure tests was that mica mat was satisfactory in this respect for application in traction motors if the silicone was kept below 15% and the cure was sufficient. Also catalysts were necessary for the cure and, if extreme resistance to pressure was needed, the resin content should be kept below 7%.

The crush resistance at 200 C today is considerably better than it was at the time when regular production was started. The silicone-treated flat sheet material which is used as a wrapper on armature coils has a crush resistance of around 10,000 psi at 200 C. The glass-backed tape has a crush resistance of around 8,000 psi at 200 C. These values are both lower than the 18,000 psi of the mica tape which they replace, but they are satisfactory for most normal use.

Life Tests

Life tests, carried out on both field and armature coils, were involved and lengthy. The results of one typical test may perhaps be of interest. This demonstrated that mica mat was a better insulation than copper was a conductor at high temperatures, and that the limiting feature in a motor might not be the insulation, but the copper. In this test a *GE-733* traction motor was used with armature coils insulated with $3^1/_4$ wrappings of glass-backed silicone sheet material. This motor armature was run in a torsional shake test stand for 250 hours at 200 C and for 90 hours cycling from 30 C to 200 C. The amplitude of vibration was approximately 0.021-inch at 27 *G*'s. At the end of this time the insulation still appeared to be in good condition, but the binding wire was loose and the wedges were charred and loose. The insulation, both the standard mica tape and the mica-mat armature insulation, was intact. The armature was then heated to 250 C and the amplitude of vibration was doubled. After 200 hours of this severe punishment the copper wires broke and short-circuited to ground, ending the test. This armature is shown in Fig. 6 as it came off the test. Note that some of the copper wires on the right-hand end windings have had their varnished-glass and mica-tape insulation completely shaken off.

Three coils from this armature are shown in Fig. 7. The two on the left are conventional coils with varnished-asbestos turn insulation and paper-backed silicone-pasted mica tape as ground insulation. The coil on the right had a mica-mat, glass-backed ground insulation wrapped over the polycoil. The ordinary varnish was completely disintegrated, leaving the asbestos a white fluffy mass. Mica-mat ground insulation still had nearly its original dielectric strength even after this tremendous heat and vibration. The conventional mica-tape coils were also satisfactory except for places in the
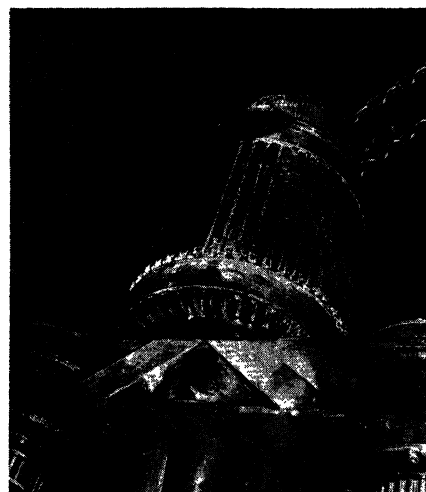
Fig. 6. Armature of 150-horsepower traction motor insulated with mica mat after thermal aging and vibration

Fig. 7. Armature coils from a heat-aging life test

armature that had come to extreme heat. Here the binder had decomposed, allowing the mica flakes to slip. These coils had failed electrically.

Field Tests

After these life tests, sample motors were put in all types of traction service ranging from trolley coaches to main line locomotives. Some of these motors are going into their fourth year of service. As the reports came in from the field and returned equipment was taken apart, it became evident that mica mat was not simply an adequate substitute for mica tape on armature coils but rather a superior insulating material. It was at once lower in cost, gave more uniformly shaped coils, would stand more heat and vibration, and was somewhat easier to use in the factory.

On the other hand, field tests showed that mica mat was not a universal substitute for mica tape. At the present time it does not seem suitable for taping armature conductors, although it makes excellent wrappers for ground insulation

on armature coils. Neither does it seem suitable for field coils where high pressures are exerted in manufacture or in use. As shown before, there are definite limitations to the pressures this material can stand. There are still other places where it has been found unsuitable because of mechanical deficiencies as compared to mica tape. Where properly engineered, it has turned out to be a most satisfactory insulating material for traction motors.

Summary

Mica mat treated with silicone resin has been developed into an excellent material for insulating traction motors. As in testing of all traction-motor insulation, the effect of heat alone was only a small part of the investigation necessary. The effects of heat in combination with humidity, with pressure, and with vibration were of equal or perhaps even greater importance than the effect of heat alone. The combination of all these effects was studied in a series of life tests and field tests of motors under operating conditions. It would be extremely difficult to classify silicone-treated mica mat on the basis of conventional $ABHC$-type heat classification. Although quite suitable at high temperatures if there are not high pressures, the material is inadequate if equipment must stand high pressures. Classification of a material like silicone-treated mica mat can only be done reasonably on the basis of use: will it stand up to the purpose for which it is designed? It is sad but true that electrical design engineers must consider each case separately on its merits. They cannot depend on all over-all heat-type classification to tell them whether a material can be used or not. In traction-motor equipment mechanical effects are extremely important. Indeed, in the case of many insulating materials tested, they have been found to be limiting factors.

References

1. Rotating Electric Machinery on Railway Locomotives and Rail Cars and Trolley, Gasoline-Electric and Oil-Electric Coaches. *AIEE Standard No. 11* and *ASA Standard No. C35.1-1943*, 1943.

2. Standard Methods of Test for Bulking Thickness of Paper. *ASTM 527–41*, American Society for Testing Materials, Philadelphia, Pa., 1941.

3. Standard Methods of Test in Tensile-Breaking Strengths of Paper and Paper Products. *ASTM D828–4, Ibid.*, 1948.

4. Methods for Chemical Analysis of Steel. *ASTM E30–52, Ibid.*, para. 6, 1952.

5. Methods of Test for Dielectric Strength of Electrical Insulating Materials at Commercial Power Frequencies. *ASTM D149, Ibid.*, 1944.

The Effect of Pole and Zero Locations on the Transient Response of Sampled-Data Systems

ELIAHU I. JURY
ASSOCIATE MEMBER AIEE

Synopsis: The effect of the location of the poles and zeros of the transfer function of a sampled-data system on the location and magnitude of the maxima and minima of transient response resulting from a step function input is studied. Theorems are given relating to the necessary conditions for the production of monotonic and non-monotonic time response expressed in terms of pole and zero locations in the z-plane. It is shown that, under certain conditions of pole and zero locations, the normalized time-sequence response may well be approximated by a single dominant time term. A method is presented of ascertaining from the pole and zero location whether these conditions exist. On the basis of dominant term approximation, the methods of synthesis applied to second-order systems can be generally extended to general systems. The results of this investigation are directly applicable to the design problems in the field of pulsed networks and sampled-data control systems.

IT IS known that the z-transform method[1-4] can be readily used for analysis and design of sampled-data systems. This method is essentially a Laplace transform technique in which the sampled output is related to the input by system functions in the form of z-transforms. These system functions are ratios of rational polynomials in the variable z where z is defined as ϵ^{Ts} and T is the sampling period.

These system functions can generally be regarded to consist of zeros and pole in the z-plane and the location of these poles and zeros has considerable bearing on the response and design of sampled-data systems. For instance, in the synthesis of a sampled-data configuration shown in Fig. 1, use of a linear network is employed either in the forward or the feedback path. The effect of these networks on the system response can best be visualized by noting the effect of pole and zero configuration of the sampled-data system in the z-plane. More recently digital computers[5] or processing units[6] which are in essence sampled-data devices have been widely employed as a means of stabilizing sampled-data systems, as shown in Fig. 2. Thus their effect on the system zero pole pattern requires special study.

In this paper an investigation is made of the effect that the location of the poles and zeros of the system-function (output/input) of a linear sampled-data system has upon the location and magnitude of the maxima and minima in the transient response of the system when a step-function input is applied. Attention is also given to the necessary conditions for the production of a monotonic time-sequence response output for the system, these conditions being expressed in terms of pole and zero location. The investigation is limited to stable systems whose transfer function contains no poles outside or on the unit circle. It is shown that, under certain conditions of pole and zero location, the normalized time response may well be approximated by a single dominant time term. When this condition exists the design procedures for the second-order system can be extended to general systems. The results of this investigation are applicable to pulsed net-

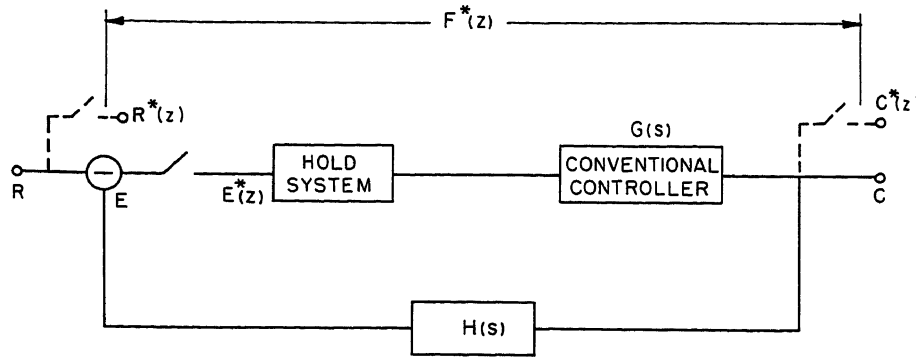**Fig. 1. Typical sampled-data control system**

works or to sampled-data systems employing networks or digital computers as a means of stabilization.

Nomenclature

The explanation of z-transforms and the time-sequence response notation used in this paper is as follows:

$\alpha_k \pm j\beta_k$ = location of kth pair of complex poles inside the unit circle

$\zeta_k \pm j\gamma_k$ = location of kth pair of complex zeros in the z-plane

ρ_k = location of kth real pole inside the unit circle

μ_k = location of kth real zero in the z-plane

$\beta_{0k}{}^2 = (1-\alpha_k)^2 + \beta_k{}^2$

$\gamma_{0k}{}^2 = (1-\zeta_k)^2 + \gamma_k{}^2$

ρ_{ik} = distance from ρ_i to $\alpha_k + j\beta_k$

μ_{ik} = distance from μ_i to $\alpha_k + j\beta_k$

β_{ika} = distance from $\alpha_i + j\beta_i$ to $\alpha_k + j\beta_k$

β_{ikb} = distance from $\alpha_i - j\beta_i$ to $\alpha_k + j\beta_k$

γ_{ika} = distance from $\zeta_i + j\gamma_i$ to $\alpha_k + j\beta_k$

γ_{ikb} = distance from $\zeta_i - j\gamma_i$ to $\alpha_k + j\beta_k$

$\bar{\rho}_{ik}$ = distance from ρ_i to $\zeta_k + j\gamma_k$

$\theta_{ik} = \theta_{ika} + \theta_{ikb} = \tan^{-1}\dfrac{\beta_k - \gamma_i}{\alpha_k - \zeta_i} +$

$$\tan^{-1}\dfrac{\beta_k + \gamma}{\alpha_k - \zeta_i}$$

$\psi_{ik} = \psi_{ika} + \psi_{ikb} = \tan^{-1}\dfrac{\beta_k - \beta_i}{\alpha_k - \alpha_i} +$

$$\tan^{-1}\dfrac{\beta_k + \beta_i}{\alpha_k - \alpha_i}$$

$\delta_{ki} = \tan^{-1}\dfrac{\beta_k}{\alpha_k - \mu_i}$

$\phi_{ik} = \tan^{-1}\dfrac{\beta_k}{\alpha_k - \rho_i}$

$\theta_k = \tan^{-1}\dfrac{\beta_k}{\alpha_k}$

$\psi_k = -\tan^{-1}\dfrac{\beta_k}{\alpha_k - 1}$

Development and Application of Theory

NORMALIZED TIME-SEQUENCE RESPONSE

To provide sufficient generality, it will be assumed that the sampled-data system under consideration has a transfer func-

tion (output/input) containing l pairs of conjugate complex poles, q real poles, g pairs of conjugate complex zeros, and m real zeros. (It is assumed that no multiple poles exist; also the real zeros can be outside the unit circle in the z-plane.) It is further assumed from physical realizability that $2l+q>2g+m$. The notations used are explained in the Nomenclature and are illustrated in Fig. 3.

If the system shown in Fig. 1 is subjected to a unit step-function input, the z-transform of the output time-sequence response is of the form

$C^*(z) = R^*(z)F^*(z) =$

$$\frac{Kz \prod\limits_{i=1}^{m}(z-\mu_i)\prod\limits_{i=1}^{g}[(z-\zeta_i)^2+\gamma_i{}^2]}{(z-1)\prod\limits_{i=1}^{q}(z-\rho_i)\prod\limits_{i=1}^{l}[(z-\alpha_i)^2+\beta_i{}^2]} \quad (1)$$

where K is a scale factor and can be assumed to be unity in the following work.

The inverse z-transformation[4] of equation 1 yields a time response sequence $C(nT)$ containing a constant term A_0 plus terms which are functions of time. For convenience in the comparison of the responses of systems having different values of A_0, it is desirable to consider the amplitude-normalized time response sequence defined as

$$C(nT) = A_0 C_0(nT) \quad (2)$$

in which

$$A_0 = \frac{\prod\limits_{i=1}^{m}(1-\mu_i)\prod\limits_{i=1}^{g}[(1-\zeta_i)^2+\gamma_i{}^2]}{\prod\limits_{i=1}^{q}(1-\rho_i)\prod\limits_{i=1}^{l}[(1-\alpha_i)^2+\beta_i{}^2]} \quad (3)$$

Accordingly, by inverse transformation[4] of equation 1 and normalization, the following expression is obtained for the normalized time-sequence response

$$C_0(nT) = 1 + \sum_{k=1}^{l} M_k \frac{\beta_{0k}{}^2}{\beta_k[(\alpha_{k-1})^2+\beta_k{}^2]^{1/2}} \times$$

$$(\alpha_k{}^2+\beta_k{}^2)^{n/2}\sin(n\theta_k+\lambda_k+\psi_k)+$$

$$\sum_{k=1}^{q}(-1)^{k+r}{}_k N_k(\rho_k)^n \quad (4)$$

in which

$$M_k = \frac{\prod\limits_{i=1}^{m}\left(\dfrac{\mu_{ik}}{1-\mu_i}\right)\prod\limits_{i=1}^{g}\left(\dfrac{\gamma_{ika}}{\gamma_{0i}}\right)\left(\dfrac{\gamma_{ikb}}{\gamma_{0i}}\right)}{\prod\limits_{i=1}^{q}\left(\dfrac{\rho_{ik}}{1-\rho_i}\right)\prod\limits_{\substack{i=1\\i\neq k}}^{l}\left(\dfrac{\beta_{ika}}{\beta_{0i}}\right)\left(\dfrac{\beta_{ikb}}{\beta_{0i}}\right)} \quad (5)$$

$$N_k = \frac{\prod\limits_{i=1}^{m}\dfrac{|\mu_i-\rho_k|}{1-\mu_i}\prod\limits_{i=1}^{g}\left(\dfrac{\bar{\rho}_{ki}}{\gamma_{0i}}\right)^2}{\prod\limits_{i=1}^{l}\left(\dfrac{\rho_{ki}}{\beta_{0i}}\right)^2\prod\limits_{\substack{i=1\\i\neq k}}^{q}\left|\dfrac{\rho_i-\rho_k}{1-\rho_i}\right|} \quad (6)$$

$$\lambda_k = \sum_{i=1}^{g}\theta_{ik} - \sum_{\substack{i=1\\i\neq k}}^{l}\psi_{ik} +$$

$$\sum_{i=1}^{m}\delta_{ik} - \sum_{i=1}^{q}\phi_{ik} \quad (7)$$

and r_k is the number of real zeros greater than ρ_k in the z-plane. It is noticed that the coefficients M_k and N_k are formed in the same manner. Each is equal to the quotient of the product of factors due to zeros divided by the product of factors due to poles. The factors for all poles and zeros have the same form.

An example using most of the important terms in equations 1 through 7 is given in Appendix I.
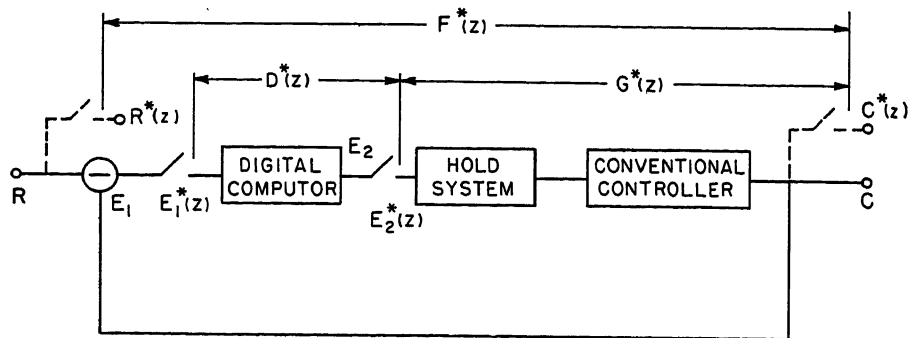

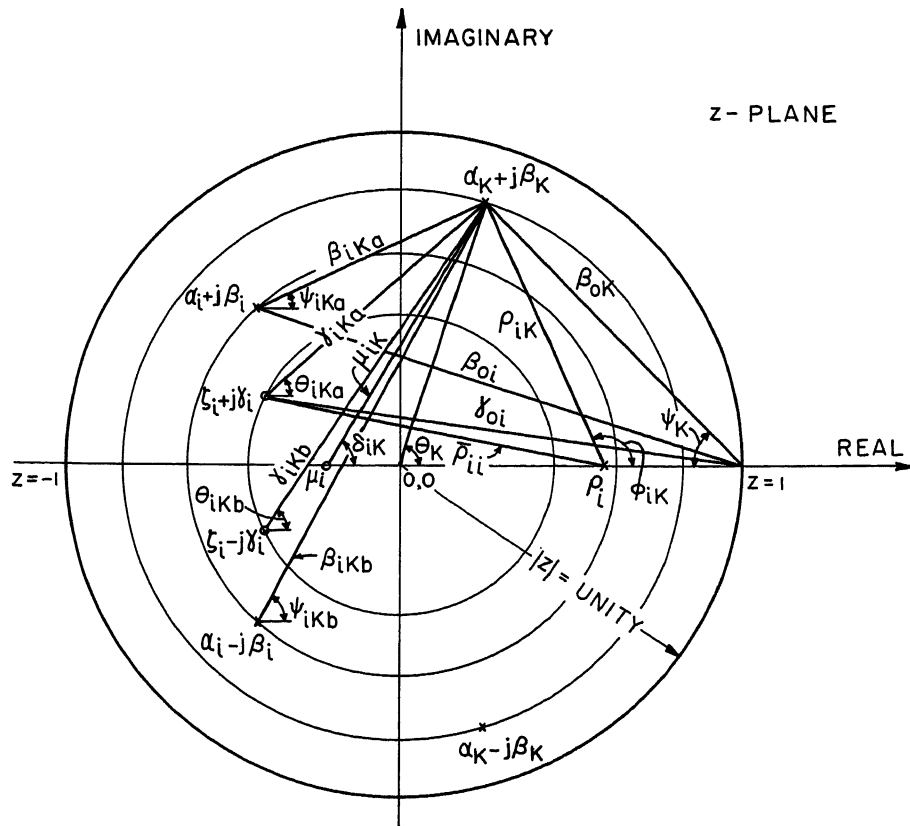**Fig. 2. Typical sampled-data control system employing a digital computer**

Fig. 3. Illustration of the notations employed

$$C^*(z) \times (z-1) = z \times$$

$$\frac{\prod\limits_{i=1}^{m}(z-\mu_i) \prod\limits_{i=1}^{g}[(z-\zeta_i)^2+\gamma_i^2]}{\prod\limits_{i=1}^{q}(z-\rho_i) \prod\limits_{i=1}^{l}[(z-\alpha_i)^2+\beta_i^2]} \quad (9)$$

Inverse z-transformation of equation 9 yields $\Delta C(nT)$, which may be written as

$$\Delta C(nT) = A_1'(\alpha_1^2+\beta_1^2)^{n/2} G(nT) \quad (10)$$

where

$$G(nT) = \sin(n\theta_1+\lambda_1) +$$

$$\sum_{k=2}^{l}\frac{\beta_1}{\beta_k}m_k'\left(\frac{\alpha_k^2+\beta^2_k}{\alpha_1^2+\beta_1^2}\right)^{n/2} \times$$

$$\sin(n\theta_k+\lambda_k) +$$

$$\sum_{k=1}^{q}(-1)^{r_k+k+1}\frac{\beta_1}{\rho_{k1}}n_k'\frac{\rho_k{}^n}{(\alpha_1^2+\beta_1^2)^{n/2}} \quad (11)$$

$$A_1' = \frac{\prod\limits_{i=1}^{m}\mu_{i1}\prod\limits_{i=1}^{g}(\gamma_{i1a}\gamma_{i1b})}{\beta_1\prod\limits_{i=1}^{q}\rho_{i1}\prod\limits_{i=2}^{l}(\beta_{i1a})(\beta_{i1b})} \quad (12)$$

$$m_k' = \frac{\prod\limits_{i=1}^{m}\left(\dfrac{\mu_{ik}}{\mu_{i1}}\right)\prod\limits_{i=1}^{g}\left(\dfrac{\gamma_{ika}}{\gamma_{i1a}}\right)\left(\dfrac{\gamma_{ikb}}{\gamma_{i1b}}\right)}{\prod\limits_{i=1}^{q}\left(\dfrac{\rho_{ik}}{\rho_{i1}}\right)\prod\limits_{\substack{i=2\\i\neq k}}^{l}\left(\dfrac{\beta_{ika}}{\beta_{i1a}}\right)\left(\dfrac{\beta_{ikb}}{\beta_{i1b}}\right)} \quad (13)$$

$$n_k' = \frac{\prod\limits_{i=1}^{m}\left|\dfrac{\mu_i-\rho_k}{\mu_{i1}}\right|\prod\limits_{i=1}^{g}\left(\dfrac{\bar{\rho}_{ki}}{\gamma_{i1a}}\right)\left(\dfrac{\bar{\rho}_{ki}}{\gamma_{i1b}}\right)}{\prod\limits_{i=2}^{l}\left(\dfrac{\rho_{ki}}{\beta_{i1a}}\right)\left(\dfrac{\rho_{ki}}{\beta_{i1b}}\right)\prod\limits_{\substack{i=1\\i\neq k}}^{q}\dfrac{|\rho_i-\rho_k|}{\rho_{i1}}} \quad (14)$$

COMPOSITION OF COEFFICIENTS AND OF PHASE ANGLES

Referring to Fig. 4 and equation 5, it is evident that the pair of complex poles at $\alpha_i\pm j\beta_i$ contributes a multiplying factor f to the denominator of M_k of the amount

$$f = \frac{r_1 r_2}{r_0^2} \quad (8)$$

In general a typical factor for a single element (pole or zero) is equal to the ratio of two distances in the z-plane, namely, the distance from the element to the pole for which the coefficients M_k and N_k are to be computed divided by the distance of the element from the edge of the unit circle (or when $z=1$).

It is possible, however, to determine how the poles $\alpha_i\pm j\beta_i$ must move in the z-plane in order that the factor f remain constant. A constant f chart for the typical pole angle ψ_k can be plotted from which the coefficient M_k can be readily obtained for the pair of poles of angle ψ_k.

To determine the effect of pole and zero locations on the typical phase angle λ_k, Fig. 4 should be examined. It is noticed that the angle contribution of a pair of complex elements $\alpha_i\pm j\beta_i$ to λ_k is equal to the angle substended by the pair of poles $\alpha_k\pm j\beta_k$ at the point $\alpha_i\pm j\beta_i$. This is the angle B indicated in the same figure.

Constant-angle contours in the z-plane can be plotted from which a particular λ_k can be easily determined. Charts of constant f and constant λ_k are not plotted in this paper. However, their plot can be easily computed if required.

TWO TIME-RESPONSE THEOREMS

As a result of the condition $2l+q>2g+m$, $\lim\limits_{n\to 0}C(nT)=0$. Therefore, the z-transform of the first difference of the time-sequence response is of the form (see Appendix II of reference 4)

Fig. 4. Construction for the determination of contribution (a) to the multiplying factor f, (b) to total phase angle λ_k

$$f = \frac{r_1\ r_2}{r_0^2}$$

$$B = \psi_{iK} = \psi_{iKa} + \psi_{iKb}$$

z- PLANE

REAL
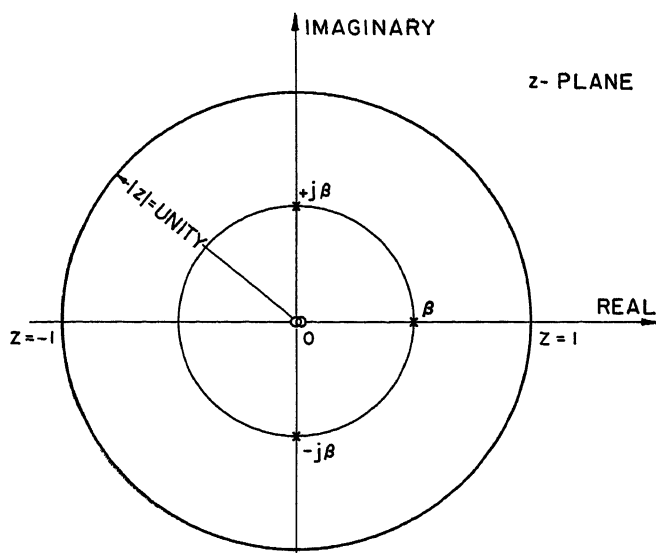
Fig. 5 (left). Pole and zero locations in the z-plane of the illustrative system
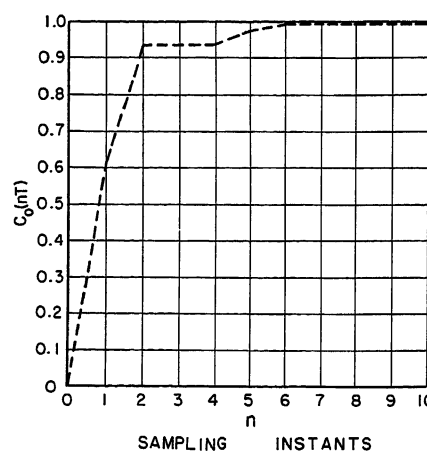
Fig. 6 (right). Transient response to a step-function input of the illustrative system

The definition of $G(nT)$ is of advantage in determining when the difference equation 10 changes sign, and hence the maxima and minima of $C(nT)$, the time-response sequence. It is evident that $\Delta C(nT)$ changes sign when $G(nT)$ changes sign, for the quantity $A_1'(\alpha_1{}^2+\beta_1{}^2)^{n/2}$ is always positive.

If it is assumed from the pole numbering that $\sqrt{\alpha_k{}^2+\beta_k{}^2}<\sqrt{\alpha_1{}^2+\beta_1{}^2}$ and if it is also assumed that $\rho_k<(\alpha_1{}^2+\beta_1{}^2)^{1/2}$, then it is evident from equation 11 that after some value of n all terms except the first in $G(nT)$ will be less than some arbitrarily small amount, and hence $G(nT)$ will change sign with the changed sign of sin $(n\theta_1+\lambda_1)$. From this behavior is deduced the following theorem:

Theorem Ia: The transient-sequence response to a step-function input of a linear sampled-data control system having no real poles between the unit circle and the circle representing the first pair of complex poles cannot be monotonic.

It is known that the z-transform method yields the response at sampling instances without any information between sampling instances. However, theorem Ia can be extended to the actual response between sampling instances, for if the sampled output cannot be monotonic it necessarily follows that the output between sampling instances is also nonmonotonic, thus:

Theorem Ib: The transient response to a step-function input of a linear sampled-sampled-data control system having no real poles between the unit circle and the circle representing the first pair of complex poles cannot be monotonic.

It is of interest to examine next the case where $\rho_1=(\alpha_1{}^2+\beta_1{}^2)^{1/2}$. Under such conditions, $G(nT)$ becomes

$$G(nT)=(-1)^{r_1+2}n_1'\frac{\beta_1}{\rho_{11}}+\sin(n\theta_1+\lambda_1)+$$

$$\sum_{k=2}^{l}\frac{\beta_1}{\beta_k}\left(\frac{\alpha_k{}^2+\beta_k{}^2}{\alpha_1{}^2+\beta_1{}^2}\right)^{n/2}m_k'\times$$

$$\sin(n\theta_k+\lambda_k)+$$

$$\sum_{k=2}^{q}(-1)^{r_k+k+1}n_k'\frac{\beta_1}{\rho_{k1}}\frac{\rho_k{}^n}{(\alpha_1{}^2+\beta_1{}^2)^{n/2}}\quad(15)$$

In order that the time-sequence response $C_0(nT)$ does not exhibit maxima nor minima, it is necessary that $\Delta C(nT)$ and hence $G(nT)$ be nonnegative. It is seen from equation 15 that as n becomes large, $G(nT)$ can be approximated well by the first two terms. Thus in order that $G(nT)$ remain nonnegative, it is necessary that r_1 be even and that $n_1' \beta_1/\rho_{11}$ be equal to or greater than the upper value of sin $(n\theta_1+\lambda_1)$. From these facts can be deduced the following theorem:

Theorem II: In order for the transient-response to a step function of a sampled-data system having its real pole on the perimeter of the same circle on which the first pair of complex poles lies to increase monotonically, it is necessary that the number of real zeros larger than the real pole in the z-plane be zero or even, and that the quantity $n_1'\beta_1/\rho_{11}$ be equal to or greater than the upper limit of sin $(n\theta_1+\lambda_1)$. In general, sin $(n\theta_1+\lambda_1)$ is not periodic and its maximum value may not necessarily be unity. However, to assure the monotonic condition it is evident that $n_1'\beta_1/\rho_{11}$ should be larger or equal to unity. In theorum II, it is assumed that the real pole is positive; however, for the negative real pole see Appendix II.

The conditions given for theorem II are necessary but not sufficient for the productions of a monotonic time-sequence response.

Illustrative Example

As an illustration of the use of the second theorem and the various equations derived, the system consisting of multiple zeros at the origin, one real pole β, and two complex poles at $\pm j\beta$ will be considered, as shown in Fig. 5. Referring to equation 15, $r_1=0$, $\beta_1/\rho_{11}=1/\sqrt{2}$, n_1' equals unity. Since in this example $\theta_1=\pi/2$, $\lambda_1=\pi/4$, the upper limit of sin $(n\theta_1+\lambda_1)=+1/\sqrt{2}$. Therefore, according to theorem II, it would be expected that the response is monotonic and exhibits no maxima or minima. The normalized time-response equation for such a system can be obtained from equation 4, thus

$$C_0(nT)=1+\frac{\beta^n(1+\beta^2)^{1/2}}{\sqrt{2}}\times$$

$$\left[(1-\beta)\sin(n\theta+\psi+\lambda)-\frac{(1+\beta^2)^{1/2}}{\sqrt{2}}\right]\quad(16)$$

Assume

$$\beta=0.5\quad(17)$$

Thus

$$\theta=\pi/2$$
$$\lambda=-3\pi/4+\pi=\pi/4\quad(18)$$
$$\psi=-\tan^{-1}\beta/-1=(26.5°-\pi)$$

Equation 16 becomes

$$C_0(nT)=1+(0.5)^n\frac{1.12}{1.42}\left[0.5\sin\left(\frac{n\pi}{2}-\pi+\right.\right.$$

$$\left.\left.26.5°+\frac{\pi}{4}\right)-0.707\times1.12\right]\quad(19)$$

The plot of equation 19 is shown in Fig. 6, where the response is monotonic and exhibits no maxima or minima. However, any zero not at the origin would make n_1' less than unity and the system would be nonmonotonic because the upper limit of sin $(n\theta+\lambda)$ in this case would be larger than $1/\sqrt{2}$.

DOMINANCE OF NORMALIZED TIME-SEQUENCE RESPONSE

From the preceding section, it appears that it is possible to approximate $\Delta C(nT)$ after a certain n, by the first term of

equation 11, i.e., $\sin(n\theta_1+\lambda_1)$. This term results from the poles $\alpha_1\pm j\beta_1$ which are on the perimeter of the circle in the z-plane nearest the unit circle. Such an approximation, when applicable, is of considerable value in simplifying the determination of the locations and magnitudes of the maxima and minima of the time-sequence response. To find the points of maxima and minima the values of $2n$ must be obtained, for which $\Delta C(nT)$ changes sign. This value of n ought to be an integer and can be found by first solving the following equation

$$n=\left(\frac{h\pi-\lambda_1}{\theta_1}\right)+Y_{nh} \tag{20}$$

where $h=$ integer for which $(h\pi-\lambda_1)$ is positive, and Y_{nh} is a corrective factor whose value is such that would let n be an integer number. (The value of n can be obtained by first solving the equation $(n\theta_1+\lambda_1)=h\pi$, which would generally yield a noninteger value, and the actual n is the upper integer number of the value obtained.) The magnitude fo this correction factor depends on the value of h used.

To consider the magnitude of $C(nT)$ at the critical points of maxima and minima, it will be convenient to consider the deviation of $C_0(nT)$ from unity at these maxima and minima. For this purpose the quantity γ_h will be introduced, being defined as the deviation of $C_0(nT)$ from unity at the critical point corresponding to the value of h used, i.e.

$$\gamma_h=C_0\left(\frac{h\pi-\lambda_1}{\theta_1}+Y_{nh}\right)-1 \tag{21}$$

By evaluating equation 4 at the critical points defined by equation 20 and applying equation 21, it is found that

$$\gamma_h=M_1\frac{\beta_{01}{}^2}{\beta_1[(\alpha_1-1)^2+\beta_1{}^2]^{1/2}}\times$$
$$(\alpha_1{}^2+\beta_1{}^2)^{\left(\frac{h\pi-\lambda_1+Y nh\theta_1}{2\theta_1}\right)}\times$$
$$\sin(h\pi+\psi_1+Y_{nh}\theta_1)+$$
$$\sum_{k=2}^{l}M_k\frac{\beta_{0k}{}^2}{\beta_k[(\alpha_k-1)^2+\beta_k{}^2]^{1/2}}\times$$
$$(\alpha_k{}^2+\beta_k{}^2)^{\left(\frac{h\pi-\lambda_1+Y nh\theta_1}{2\theta_1}\right)}\times$$
$$\sin[\theta_k\frac{h\pi-\lambda_1+Y nh\theta_1}{\theta_1}+\lambda_k+\psi_k]+$$
$$\sum_{K=1}^{q}(-1)^{rk+k}N_k\rho_k{}^{\left(\frac{h\pi-\lambda_1+Y nh\theta_1}{\theta_1}\right)} \tag{22}$$

It is desirable to introduce two factors which will serve as measures of the dominance γ_h by the first term of equation 22. These are k_{Mkh}, and k_{Nkh}, and are defined as follows

$$k_{Mkh}=\frac{M_1\frac{\beta_{01}{}^2}{\beta_1[(\alpha_1-1)^2+\beta_1{}^2]^{1/2}}\times(\alpha_1{}^2+\beta_1{}^2)^{\left(\frac{h\pi-\lambda_1+Y nh\theta_1}{2\theta_1}\right)}\times\sin(h\pi+\psi_1+Y_{nh}\theta_1)}{M_k\frac{\beta_{01}{}^2}{[(\alpha_k-1)^2+\beta_k{}^2]^{1/2}}\times(\alpha_k{}^2+\beta_k{}^2)^{\left(\frac{h\pi-\lambda_1+Y nh}{2\theta_1}\right)}} \tag{23}$$

$$k_{Nkh}=\frac{M_1\frac{\beta_{01}{}^2}{\beta_1[(\alpha_1-1)^2+\beta_1{}^2]^{1/2}}\times(\alpha_1{}^2+\beta_1{}^2)^{\left(\frac{h\pi-\lambda_1+Y nh\theta_1}{2\theta_1}\right)}\times\sin(h\pi+\psi_1+Y_{nh}\theta_1)}{N_k\rho_k{}^{\left(\frac{h\pi-\lambda_1+Y nh\theta_1}{\theta_1}\right)}} \tag{24}$$

In terms of these new factors, equation 22 can be written as

$$\gamma_h=\gamma_1\left\{1+\sum_{k=2}^{l}k_{Mkh}\times\right.$$
$$\sin[\theta_k\frac{h\pi-\lambda_1+Y nh\theta_1}{\theta_1}+\lambda_k+\psi_k]+$$
$$\left.\sum_{K=2}^{q}(-1)^{rk+k}\frac{1}{k_{Nkh}}\right\} \tag{25}$$

where γ_1 is the first term of equation 22.

The importance of the factors k_{Mkh} and k_{Nkh} is now apparent. If these quantities are sufficiently large, then the contribution resulting from terms other than the first term in equation 25 will cause a small correction, resulting in a dominance of the deviation by the first term. It is seen from equations 23 and 24 that k_{Mkh} and k_{Nkh} are large if:

1. $\alpha_1{}^2+\beta_1{}^2$ is large, i.e., the greater the distance of the first pair of complex poles from the origin the greater is k_{MKh}.

2. $\alpha_k{}^2+\beta_k{}^2$ are small, i.e., all other complex poles exist near the origin of the unit circle.

3. ρ_k is small, i.e., all real poles are near the origin in the z-plane. Evidently when the foregoing conditions exist, $C_0(nT)$ can be approximated in equation 26, where it is assumed that

$$\frac{M_1}{\beta_1}>>\frac{M_k}{\beta_k} \text{ and } M_1\frac{\beta_{01}{}^2}{\beta_1[(\alpha_1-1)^2+\beta_1{}^2]^{1/2}}>>N_k$$

$$C_0(nT)=1+M_1\frac{\beta_{01}{}^2}{\beta_1[(\alpha_1-1)^2+\beta_1{}^2]^{1/2}}\times$$
$$(\alpha_1{}^2+\beta_1{}^2)^{n/2}\sin(n\theta_1+\lambda_1+\psi_1) \tag{26}$$

where

$$M_1=\frac{\prod_{i=1}^{m}\left(\frac{\mu_{i1}}{1-\mu_i}\right)\prod_{i=1}^{q}\left(\frac{\gamma_{i1a}}{\gamma_{0i}}\right)\left(\frac{\gamma_{i1b}}{\gamma_{0i}}\right)}{\prod_{i=1}^{q}\left(\frac{\rho_{i1}}{1-\rho_i}\right)\prod_{i=2}^{l}\left(\frac{\beta_{i1a}}{\beta_{0i}}\right)\left(\frac{\beta_{i1b}}{\beta_{0i}}\right)} \tag{27}$$

$$\lambda_1=\sum_{i=1}^{q}\theta_{i1}-\sum_{i=2}^{l}\psi_{i1}+\sum_{i=1}^{m}\delta_{i1}-\sum_{i=1}^{q}\phi_{i1} \tag{28}$$

From the foregoing discussion it appears that, when dominance holds, the first overshoot for $h=1$ can be obtained from equation 22.

$$\gamma_1=M_1\frac{\beta_{01}{}^2}{\beta_1[(\alpha_1-1)^2+\beta_1{}^2]^{1/2}}\times$$
$$(\alpha_1{}^2+\beta_1{}^2)^{\left(\frac{\pi-\lambda_1+Y n1\theta_1}{2\theta_1}\right)}\times$$
$$\sin(\pi+\psi_1+Y_{n1}\theta_1) \tag{29}$$

Equation 29 shows that the deviation γ_1 depends on the quantity

$$\left(\frac{\beta_{01}{}^2}{\beta_1[(\alpha_1-1)^2+\beta_1{}^2]^{1/2}}\right)\times$$
$$(\alpha_1{}^2+\beta_1{}^2)^{\left(\frac{\pi-\lambda_1+Y n1\theta_1}{2\theta_1}\right)}\times$$
$$\sin(\pi+\psi_1+Y_{n1}\theta_1)$$

multiplied by a factor M_1, due to the location of the poles and zeros in the system other than the first pair of poles. It can easily be shown that the deviation γ from unity at the first maximum of the normalized time-sequence response resulting from the application of a step-function input to a second-order system (i.e., consisting of one pair of complex poles and a single real zero in the z-plane) is as follows

$$\gamma=\frac{\beta_{01}{}^2(\mu_{i1}/1-\mu_i)}{\beta_1[(\alpha_1-1)^2+\beta_1{}^2]^{1/2}}\times$$
$$(\alpha_1{}^2+\beta_1{}^2)^{\left(\frac{\pi-\lambda+Y n\theta_1}{2\theta_1}\right)}\times$$
$$\sin(\pi+\psi_1+Y_n\theta_1) \tag{30}$$

Thus the result shown in the foregoing is that, except for a difference between λ_1 and λ in equations 29 and 30, the deviation from unity γ_1 is that of the system containing only the dominant pair of poles $(\alpha_1\pm j\beta_1)$ multiplied by a factor due to every other pole and zero (excluding μ_1, the zero of the second-order system) in the system. Therefore, the design information obtained for a second-order system[4] can be extended to general systems when the dominance is applicable. The effect of other poles and zeros on overshoot and rise time of a second-order system can be evaluated easily. Furthermore, the methods of analysis developed in this paper should be extremely useful in determining the locations and magnitude of the maxima and minima of C_0 (nT) in a perfectly general analysis problem even when dominance does not hold. An example illustrating the dominance of the normalized time-sequence response is presented in Appendix III.

Conclusions

The use of the z-transform method in the analysis and design of sampled-data control systems is becoming increasingly

important, because this method yields considerable information in a systematic manner of a sampled-data system's behavior. From the material presented in this paper, it is evident that the transient response to a step-function input can be obtained from pole and zero patterns of the sampled-data system plotted in a z-plane. By the application of theorems Ia, Ib and II, it is often possible to determine whether a nonmonotonic time-sequence response exists.

The analytical procedures that have been outlined allow the determination of the locations and magnitudes of the maxima and the minima to any desired degree of approximation in a direct and relatively simple manner. The calculation of the factors k_{Mkh} and k_{Nkh} serves to indicate the importance of the various terms in equation 25.

It has been shown that, if sufficient separations exist between the poles and zeros in the z-plane, i.e., when conditions 1, 2, and 3 exist, then the time response can be well approximated at the first maxima and beyond by using a simple time term. When the dominant-term theory is applicable, computation of the location and magnitude of the maxima and minima is simplified considerably. In addition, the effect of the various poles and zeros on the location and magnitude of this maximum time response can be quantitatively prescribed which would indicate the deviation from the maxima of a second-order system. This investigation will be discussed in another paper where it will be shown how to obtain the direct transfer function $G^*(z)$ from the poles and zeros of the over-all transfer function.

Finally, certain of the methods presented in this paper in addition to their use in analysis can also be applied in the solution of design problems in sampled-data systems to yield prescribed time-sequence responses to step-function inputs.

Appendix I. Illustrative Example

To illustrate the various notations introduced in equations 1 through 7, the following example is chosen which includes most of the important terms in those equations.

Consider a sampled-data system transfer function to consist of real zeros at $(-2, 0)$, $(0.8, 0)$, two complex zeros at $(0.8 \pm j1)$, one real pole at $(0.6, 0)$, and four complex poles at $(0.5 \pm j0.7)$, $(0.2 \pm j0.1)$, in the z-plane.

The z-transform of the response to a step function is of the form

$$C^*(z) = \frac{z}{z-1} \frac{(z+2)(z+0.8)[(z-0.8)^2+0.1^2]}{(z-0.6)[(z-0.5)^2+0.7^2]}$$
$$[(z-0.2)^2+1^2] \quad (31)$$

From equation 3, A_0 equals

$$A_0 = \frac{(1+2)(1-0.8)[(1-0.8)^2+1^2]}{(1-0.6)[(1-0.5)^2+0.7^2]}$$
$$[(1-0.2)^2+0.1^2] = 3.25 \quad (32)$$

The normalized time-sequence response can be written as follows

$$C_0(nT) = 1 + M_1 \frac{\beta_{01}^2}{\beta_1[(\alpha_1-1)^2+\beta_1^2]^{1/2}} \times$$
$$(\alpha_1^2+\beta_1^2)^{n/2} \sin(n\theta_1+\lambda_1+\psi_1) +$$
$$M_2 \frac{\beta_{02}^2(\alpha_2^2+\beta_2^2)^{n/2}}{\beta_2[(\alpha_2-1)^2+\beta_2^2]^{1/2}} \times$$
$$\sin(n\theta_2+\lambda_2+\psi_2) +$$
$$(-1)^{1+r_1}N_1(\rho_1)^n \quad (33)$$

M_1, M_2, N_1, λ_1, and λ_2 can be obtained from equations 5, 6, and 7 as follows

$$M_1 = \frac{\left(\frac{\mu_{11}}{1-\mu_1}\right)\left(\frac{\mu_{21}}{1-\mu_2}\right)\left(\frac{\gamma_{11a}}{\gamma_{01}}\right)\left(\frac{\gamma_{11b}}{\gamma_{01}}\right)}{\left(\frac{\rho_{11}}{1-\rho_1}\right)\left(\frac{\beta_{21a}}{\beta_{02}}\right)\left(\frac{\beta_{21b}}{\beta_{02}}\right)} \quad (34)$$

$$M_2 = \frac{\left(\frac{\mu_{12}}{1-\mu_1}\right)\left(\frac{\mu_{22}}{1-\mu_2}\right)\left(\frac{\gamma_{12a}}{\gamma_{01}}\right)\left(\frac{\gamma_{12b}}{\gamma_{01}}\right)}{\left(\frac{\rho_{12}}{1-\rho_1}\right)\left(\frac{\beta_{12a}}{\gamma_{01}}\right)\left(\frac{\beta_{12b}}{\beta_{01}}\right)} \quad (35)$$

$$N_1 = \frac{\frac{|\mu_1-\rho_1|}{1-\mu_1}\frac{|\mu_2-\rho_1|}{1-\mu_2}\left(\frac{\tilde{\rho}_{11}}{\gamma_{01}}\right)^2}{\left(\frac{\rho_{11}}{\beta_{01}}\right)^2\left(\frac{\rho_{12}}{\beta_{02}}\right)^2} \quad (36)$$

$$\lambda_1 = \theta_{11}-\psi_{21}+\delta_{11}+\delta_{21}-\phi_{11} \quad (37)$$

$$\lambda_2 = \theta_{12}-\psi_{12}+\delta_{12}+\delta_{22}-\phi_{12} \quad (38)$$

In this example

$$\mu_1 = -2, \ \mu_2 = 0.8, \ r_1 = 1$$
$$\rho_1 = 0.6, \ \zeta_1 \pm j\gamma_1 = 0.8 \pm j1$$
$$\alpha_1 \pm j\beta_1 = 0.5 \pm j0.7, \ \alpha_2 \pm j\beta_2 = 0.2 \pm j0.1 \quad (39)$$
$$\theta_1 = 0.95, \ \theta_2 = 0.46$$
$$\psi_1 = 0.95+\pi, \ \psi_2 = 0.124+\pi$$

Substituting these values in equations 34 through 38 and identifying the various terms from the Nomenclature, there results

$$M_1 = 1.5, \ M_2 = 3.64, \ N_1 = 4.9,$$
$$\lambda_1 = -2.39, \ \lambda_2 = 0.14 \quad (40)$$

Thus equation 33 finally equals

$$C_0(nT) = 1 + 1.85(0.74)^{n/2}$$
$$\sin(0.95n-1.442+\pi) + 29.5(0.05)^{n/2}$$
$$\sin(0.464n+0.265+\pi) + 4.9(0.6)^n \quad (41)$$

The normalized response is shown in Fig. 7 where it is seen that it is nonmonotonic, which is consistent with theorem Ia because it has no real pole between the unit circle and the circle presenting the first pair of complex poles.

Appendix II. Illustrative Example

To illustrate theorem II, two cases are examined.

Case I

Assume a sampled-data system transfer function consisting of a zero at the origin, a zero at the point $(-0.2, 0)$, a real pole of (0.5), and two complex poles at $(\pm j0.5)$ in the z-plane. The z-transform of the response to a step function can be written using the notations of equation 1, in the following manner

$$C^*(z) = \frac{z}{z-1} \frac{(z-\mu_1)(z-\mu_2)}{(z-\rho_1)[(z-\alpha_1)^2+\beta_1^2]} \quad (42)$$

In this case

$$\mu_1 = 0, \ \mu_2 = -0.2$$
$$\rho_1 = 0.5 \quad (43)$$
$$\alpha_1 = 0, \ \pm i\beta_1 = \pm i0.5$$

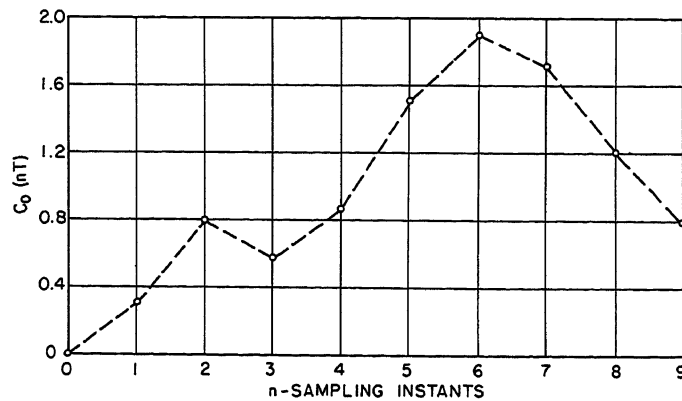Substitute relation 43 in equation 42 to obtain

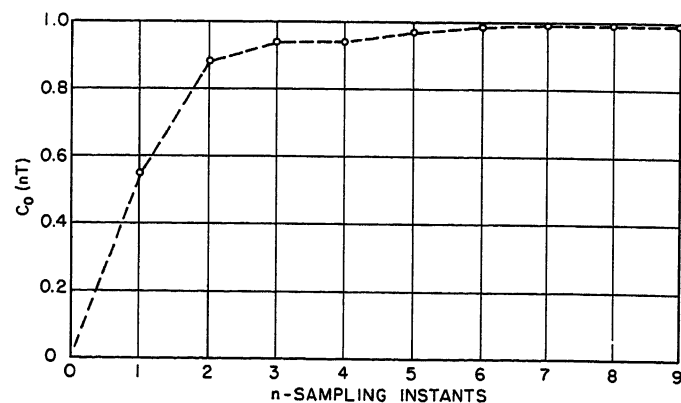Fig. 7. Transient response to a step-function input of system described in Appendix I

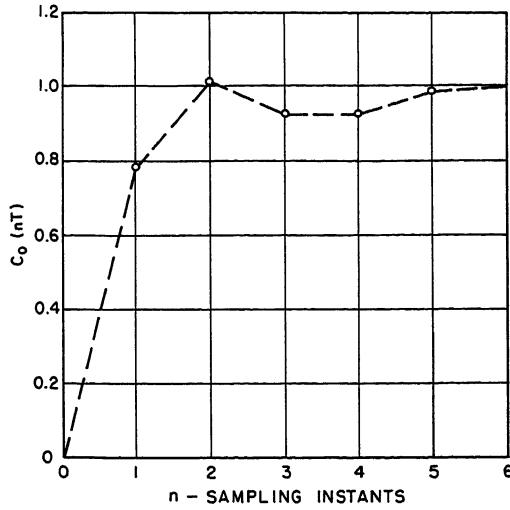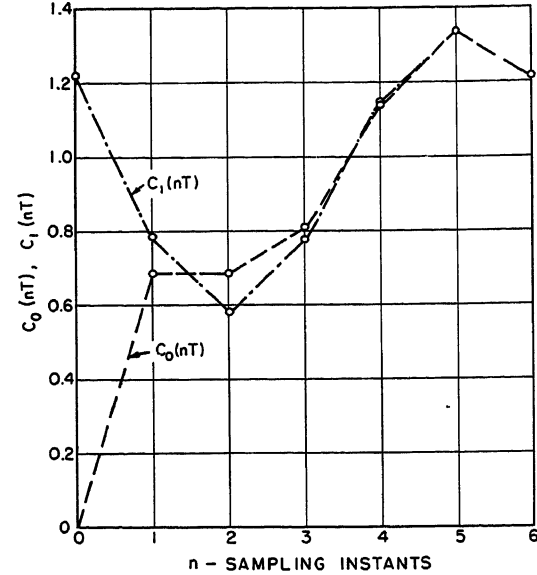Fig. 8. Transient response to a step-function input of system having monotonic response

Fig. 9 (left). Transient response of a system having non-monotonic response to a step input

Fig. 10 (right). Transient response of a system in which the dominant-term approximation is satisfactory at the first maximum and beyond

$$C^*(z) = \frac{z}{z-1} \frac{z(z+0.2)}{(z-0.5)(z^2+0.5^2)} \qquad (44)$$

The normalized system-response $C_0(nT)$, following equation 4, is given

$$C_0(nT) = 1 + M_1 \frac{\beta_{01}^2}{\beta_1[(\alpha_1-1)^2+\beta_1^2]^{1/2}}$$
$$(\alpha_1^2+\beta_1^2)^{n/2} \sin(n\theta_1+\lambda_1+\psi_1) +$$
$$(-1)^{1+r_1} N_1(\rho_1)^n \qquad (45)$$

r_1 in this case equals zero (no real zero larger than ρ_1).

The expression for M_1, N_1, and λ_1 can be obtained from equations 5, 6, and 7 as follows

$$M_1 = \frac{\dfrac{\mu_{11}}{1-\mu_1}\dfrac{\mu_{21}}{1-\mu_2}}{\dfrac{\rho_{11}}{1-\rho_1}} = \frac{\dfrac{0.5}{1}\dfrac{\sqrt{(-0.2)^2+0.5^2}}{1+0.2}}{\dfrac{\sqrt{0.5^2+0.5^2}}{1-0.5}}$$
$$= 0.159 \qquad (46)$$

$$N_1 = \frac{\dfrac{|\mu_1-\rho_1|}{1-\mu_1}\dfrac{|\mu_2-\rho_1|}{1-\mu_2}}{\left(\dfrac{\rho_{11}}{\beta_{01}}\right)^2} = \frac{\dfrac{|0-0.5|}{1}\dfrac{|-0.2-0.5|}{1+0.2}}{\dfrac{0.5^2+0.5^2}{1^2+0.5^2}}$$
$$= 0.728 \qquad (47)$$

$$\lambda_1 = \delta_{11}+\delta_{21}-\phi_{11} \qquad (48)$$

From the Nomenclature

$$\theta_1 = \tan^{-1}\frac{0.5}{0} = \frac{\pi}{2}. \qquad (49)$$

$$\psi_1 = -\tan^{-1}\frac{0.5}{1.0} = 0.46+\pi \qquad (50)$$

Therefore, the normalized time-sequence response equals

$$C_0(nT) = 1 + 0.159 \frac{1^2+0.5^2}{0.5\sqrt{1+0.5^2}}$$
$$(0.5)^n \sin(1.57n+4.01) - 0.728(0.5)^n \qquad (51)$$

The response $C_0(nT)$ is plotted in Fig. 8, where it is observed that it is monotonic. From theorem II, for the response to be monotonic, $n_1'\beta_1/\rho_{11}$ should be larger than the upper limit of $\sin(n\theta_1+\lambda_1)$. n_1' can be obtained from equation 14 as follows

$$n_1' = \frac{|\mu_1-\rho_1|}{\mu_{11}}\frac{|\mu_2-\rho_1|}{\mu_{21}} = \frac{|-0.5|}{0.5}\frac{|-0.2-0.5|}{\sqrt{0.2^2+0.5^2}}$$
$$= 1.3 \qquad (52)$$

Furthermore, the upper limit of $\sin(n\theta_1+\lambda_1)$ is 0.763. Since β_1/ρ_{11} is 0.707, thus $n_1'\beta_1/\rho_{11}$ > upper limit of $\sin(n\theta_1+\lambda_1)$, and the response is monotonic as found from equation 51.

Case II

Considering the same system as before, except that the real zero in this case is $(0.2, 0)$ in the z-plane, the z-transform of the output is given as

$$C^*(z) = \frac{z}{z-1} \frac{z(z-0.2)}{(z-0.5)(z^2+0.5^2)} \qquad (53)$$

and the normalized response is given as

$$C_0(nT) = 1 + 0.533(0.5)^n$$
$$\sin(1.57n+1.63+\pi) - 0.469(0.5)^n \qquad (54)$$

Plotting equation 54 for values of n indicates that the system response is nonmonotonic, as shown in Fig. 9.

From theorem II, for the response to be nonmonotonic, $n_1'\beta_1/\rho_{11}$ should be less than the upper limit of $\sin(n\theta_1+\lambda_1)$. In this case

$$n_1' = \frac{|\mu_1-\rho_1|}{\mu_{11}}\frac{|\mu_2-\rho_1|}{\lambda_{21}}$$
$$= \frac{|-0.5|}{\sqrt{0.5}}\frac{|0.2-0.5|}{\sqrt{0.2^2+0.5^2}} = 0.557 \qquad (55)$$

and

$$\beta_1/\rho_{11} = 1/\sqrt{2} \qquad (56)$$

The upper limit of $\sin(n\theta_1+\lambda_1)$ is 0.919. Thus $n_1'\beta_1/\rho_{11} < \sin(n\theta_1+\lambda_1)$ and the response is nonmonotonic as shown in Fig. 9.

Therefore, it is seen that when one of the zeros is positive the response is nonmonotonic, and when one of the zeros is negative the response is monotonic, which is expected from theorem II.

In theorem II it is tacitly assumed that ρ_1 is positive and equal to $\sqrt{\alpha_1^2+\beta_1^2}$; however, in case ρ_1 is negative but $|\rho_1| = \sqrt{\alpha_1^2+\beta_1^2}$, then theorem II can be easily modified, such that

$$(-1)^n n_1'\beta_1/\rho_{11}+\sin(n\theta_1+\lambda_1) > 0,$$
$$\text{for } n\text{-integer} \qquad (57)$$

Equation 57 yields the following two con-

ditions

$$n_1'\beta_1/\rho_{11} > \text{upper limit of } \sin(n\theta_1+\lambda_1)$$
$$\text{when } n \text{ is even} \qquad (58)$$

$$\sin(n\theta_1+\lambda_1) > n_1'\beta_1/\rho_{11} \text{ when } n \text{ is odd} \qquad (59)$$

Appendix III. Illustrative Example

To illustrate the dominance of the normalized time-sequence response, consider a system consisting of a real zero at the origin, two complex zeros at $(0.6\pm j1)$, and four complex poles at $(0.5\pm j0.8)$, $(0.1\pm j0.3)$. In this example the magnitude of one of the complex poles is larger than the other. Thus the z-transform of the output sequence is

$$C^*(z) = \frac{z}{z-1} \frac{z[(z-0.6)^2+1^2]}{[(z-0.5)^2+0.8^2]}$$
$$[(z-0.1)^2+0.3^2] \qquad (60)$$

The normalized time-sequence response is

$$C_0(nT) = 1 + M_1 \frac{\beta_{01}^2}{\beta_1[(\alpha_1-1)^2+\beta_1^2]^{1/2}} \times$$
$$(\alpha_1^2+\beta_1^2)^{n/2} \sin(n\theta_1+\lambda_1+\psi_1) +$$
$$M_2 \frac{\beta_{02}^2}{\beta_2[(\alpha_2-1)^2+\beta_2^2]^{1/2}} \times$$
$$(\alpha_2^2+\beta_2^2)^{n/2} \sin(n\theta_2+\lambda_2+\psi_2) \qquad (61)$$

Substituting the various values of this example in equation 61 to obtain for the normalized time-sequence response

$$C_0(nT) = 1 + 0.464(0.89)^{n/2} \sin$$
$$(1.01n-0.502+\pi) +$$
$$1.23 (0.1)^{n/2} \sin(1.25n+1.63+\pi)$$
$$\qquad (62)$$

The response $C_0(nT)$ is plotted in Fig. 10 for various values of n. It is observed that as n increases the contribution of the last term of $C_0(nT)$ is negligibly small compared to the second term which is predominant as n increases.

To obtain the value of n which yields the maximum of the time-sequence response, solve the following equation

$$n\theta_1 + \lambda_1 = h\pi \qquad (63)$$

In this example, $\theta_1 = 1.01$, $\lambda_1 = -1.51$; thus for $h = 0$

$$n = \frac{-\lambda_1}{\theta_1} = \frac{1.51}{1.01} = 1.5 \qquad (64)$$

The actual value of n (which should be an integer) is 2 and γ_{no} in this case will be 0.5. This value of n yields the minimum of the first two terms, and to obtain the maximum solve for $h = 1$.

$$n\theta_1 + \lambda_1 = \pi \qquad (65)$$

$$n = \frac{\pi - \lambda_1}{\theta_1} = \frac{4.66}{1.01} = 4.59 \qquad (66)$$

The value of n chosen is 5 and Y_{n1} in this case equals 0.41. For this value of n the last term of equation 62 is negligibly small compared to the second, and consequently dominance holds for the first maximum and beyond. Furthermore, it can be seen from the plot of Fig. 10 that $C_1(nT)$ (the first two terms of equation 62) is considerably differ-

ent than the total response, which indicates that dominance does not hold for small values of n.

The maximum overshoot can be found using equation 29 to yield the following

$$\gamma_1 = 0.464 \, (0.89)^{2.5} \sin (0.46 + \pi)$$
$$= 0.342 \qquad (67)$$

Pole and Zero Numbering

The poles and zeros are numbered consecutively according to magnitude. For instance, the real pole nearest the point (1, 0) on the edge of the unit circle is ρ_1, the next nearest is ρ_2, etc. The pair of conjugate poles lying on a circle nearest the unit circle is $\alpha_1^2 \pm j\beta_1$, the next nearest is $\alpha_2 \pm j\beta_2$, etc. Similar numbering applies to the zeros.

References

1. THEORY OF SERVOMECHANISMS, H. M. James, N. B. Nichols, R. S. Phillips. McGraw-Hill Book Company, Inc., New York, N. Y., vol. 25, chap. 6, 1947.

2. A GENERAL THEORY OF SAMPLING SERVO-SYSTEMS, D. F. Lawden. Proceedings, Institution of Electrical Engineers, London, England, vol. 98, pt. IV, Oct. 1951, pp. 31–36.

3. THE ANALYSIS OF SAMPLED-DATA SYSTEMS, J. R. Ragazzini, L. A. Zadeh. AIEE Transactions, vol. 71, pt. II, Nov. 1952, pp. 225–34.

4. ANALYSIS AND SYNTHESIS OF SAMPLED-DATA CONTROL SYSTEMS, Eliahu I. Jury. AIEE Transactions, vol. 73, pt. I, Sept. 1954, pp. 332–46.

5. ANALYSIS OF CONTROL SYSTEMS INVOLVING DIGITAL COMPUTERS, W. K. Linvill, J. M. Salzer. Proceedings, Institute of Radio Engineers, New York, N. Y., vol. 41, no. 7, July 1953, pp. 901–06.

6. SAMPLED-DATA PROCESSING TECHNIQUES FOR FEEDBACK CONTROL SYSTEMS, Arthur R. Bergen, John R. Ragazzini. AIEE Transactions, vol. 73, pt. II, Nov. 1954, pp. 236–47.

7. ANALYSIS AND SYNTHESIS OF SAMPLED-DATA CONTROL SYSTEMS, E. I. Jury. Technical Report T-1/B, Columbia University, New York, N. Y., Oct. 1953.

8. THE EFFECT OF POLE AND ZERO LOCATIONS ON THE TRANSIENT RESPONSE OF LINEAR DYNAMIC SYSTEMS, John Mulligan, Jr. Proceedings, Institute of Radio Engineers, New York, N. Y., May 1949.

Application of Germanium Power Rectifiers

R. M. CRENSHAW
ASSOCIATE MEMBER AIEE

ALTHOUGH the metal germanium is not a recent discovery, the bulk of research and development on this material has been done in the recent past. So it can be said that the discovery of the important uses of germanium has been made during the past decade. Anything which promises a revolutionary change in the field of electronics immediately becomes front-page news because electronic equipment now plays such a vital part in our industrial development as well as in our private lives. Such wide publicity has resulted in many people using the terms in a conversational way without realizing the exact meaning or the wide range of possible applications.

Several large companies have devoted much time and money to research on the possible uses of germanium. As some of this development work progressed in the research laboratories of the General

Electric Company, it was only natural that some of the engineers should begin to think of the possibilities of this metal in a direction other than electronic applications. Consequently, one branch of the development was directed toward power rectifiers. In due course a practical germanium power rectifier element was developed and has now been applied to industry in the form of a d-c power supply.

For many years the well-known copper oxide and selenium rectifiers have been important where d-c power is needed in moderate quantities for industrial plants. Now the germanium power rectifier takes its place beside the others and promises to replace them in many applications.

Germanium Power Diodes

The first look at this new power rectifier is startling because of its small physical size compared to other metallic rectifiers; see Fig. 1. Germanium is not a plentiful metal and its present cost is quite high. Fortunately a very small quantity of the metal will perform a large rectifying job. Germanium is a gray-colored crystalline metal, hard and brittle,

with a melting point slightly below 1,000 degrees centigrade. One source of the metal is a by-product of the silver and zinc mining industry. It must be carefully refined so that the final product is in single crystal form. The pure single crystal material is produced as a long cylinder, shaped like a cigar, with the diameter being controlled by the speed and temperature of processing. The material is then sliced into wafers about 0.020 inch thick and about the diameter of a dime. But this wafer does not yet have all of the required qualities of a rectifier. Although a complete chemical explanation of the atomic structure required to produce rectifying properties is quite involved, the designation "P"-type and "N"-type germanium is currently being used to describe the structure. This simply means that some impurity or unbalance of the atomic struc-
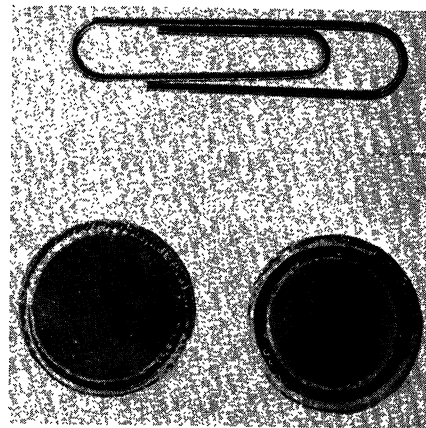

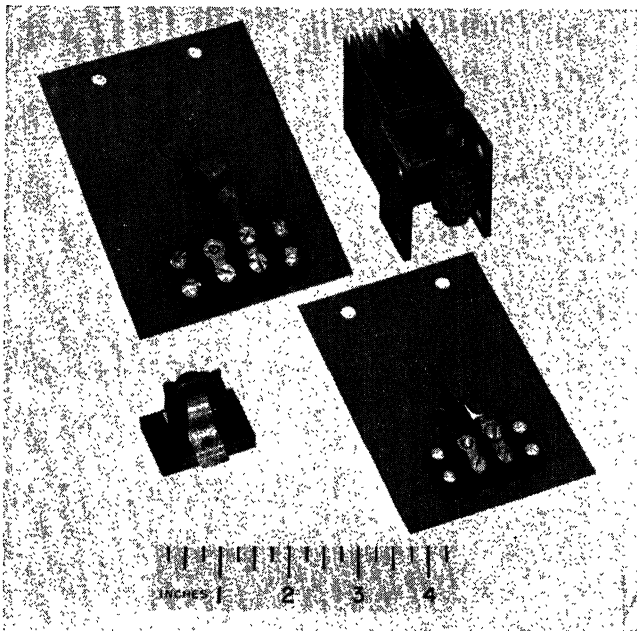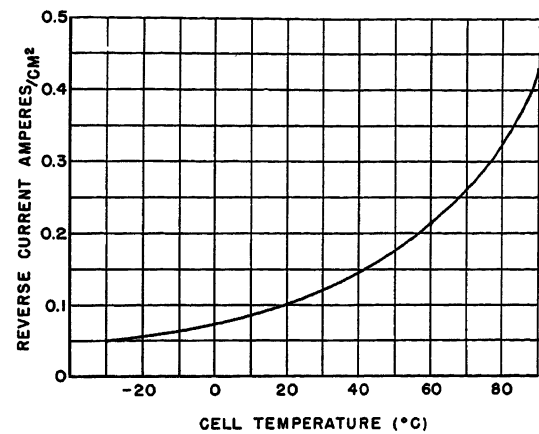Fig. 1. Germanium rectifier element

Fig. 2 (left). Germanium rectifier elements mounted

Fig. 4 (right). Reverse current–temperature characteristic at rated voltage

ture is necessary to produce free electrons. Experiment has proved that alloying with indium produces P-type germanium and alloying with antimony produces N-type germanium. The actual rectifying element is made in the form of a sandwich with the thin wafer of germanium placed between a layer of indium and a layer of antimony. This sandwich must be carefully heat-treated to produce the desired atomic structure which permits easy current flow (in the conventional sense opposite to electron flow) from P to N. This is a critical step in the process; close control of time and temperature is essential for uniformity of electrical characteristics.

Rectifier Characteristics

Fig. 2 gives some idea of the relative size and physical arrangement of the dime-sized wafers already described. The small physical size will be recognized as a distinct advantage and it is interest-ing to make a comparison between this cell and those of other metallic rectifiers.

The ratio of forward to reverse resistance is in the order of 1 to 400,000 at the area of optimum rating, which far exceeds the ratio of other metallic rectifiers. Obviously the small size is advantageous only if the material can be worked at a high current density. Fig. 3 indicates the relative current density of the three principal metallic rectifiers: germanium, copper oxide, and selenium. Note that for germanium the scale is read in amperes per square centimeter, while for the others it is milli-amperes per square centimeter. This illustrates that the germanium cell is operated at a current density almost 1,000 times greater than copper oxide. To make a more simple comparison, one of these dime-sized wafers, when properly cooled, is capable of rectifying as much as 3 kw. At this point it is evident that the small size offers some problems as well as advantages, mainly in the ability to make the physical connections which will

handle relatively large currents. Another problem is that of effectively removing the heat generated in this very small device.

Like other metallic rectifiers, germanium has a negative temperature characteristic. Forward losses are a function of load current and reverse losses are a function of inverse voltage. Both losses cause a cell temperature rise which, in turn, changes cell resistance to further modify the losses. Reverse losses increase rapidly as the temperature rises because the reverse current change is exponential. Fig. 4 indicates how reverse current changes with temperature. The heat must be dissipated from a small area very rapidly and the radiating fins with forced air or liquid immersion have proved the most effective method so far.

The normal full load temperature rise of the cell is 30 degrees centigrade over a 35-degree-centigrade ambient. Experiment has determined the derating curve for higher ambient temperature. The derating can be done either by current or voltage. Fig. 5 shows this relationship.

Electric Connections

Germanium rectifier cells can be connected for single-way or double-way,
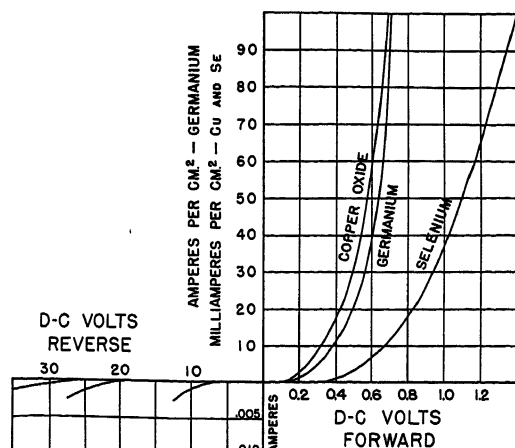
Fig. 3 (left). D-c characteristics of typical metallic rectifiers per 1-square-centimeter active area

Fig. 5 (right). Current rating at various ambient temperatures

CURRENT DERATING AT RATED VOLTAGE

PERMISSIBLE OPERATING VOLTAGE OF NOMINAL 65 VOLT RMS CELL AT VARIOUS TEMPERATURES WHEN OPERATED AT RATED CURRENT
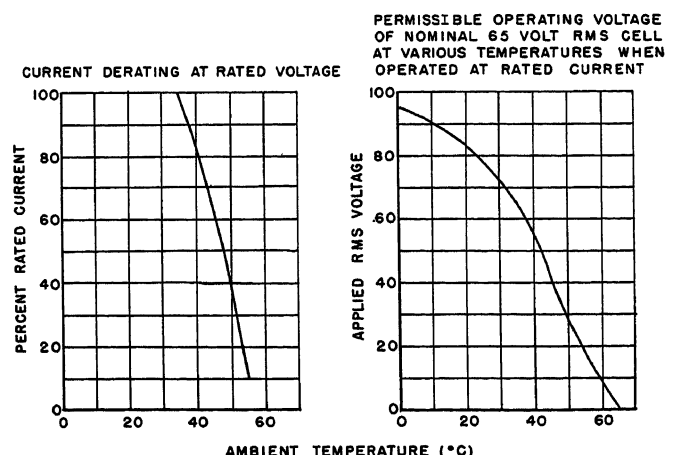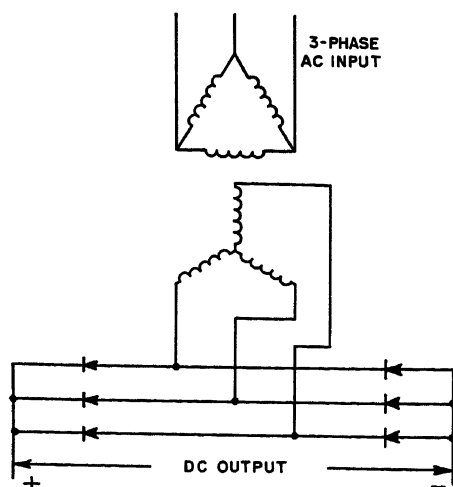
3-PHASE
AC INPUT

DC OUTPUT
+   −

single-phase and 3-phase operation. Individual cells can be operated at a voltage considerably higher than is customary with other metallic rectifiers of equivalent rating. One of the standard connections is shown in Fig. 6. This is a standard double-way, 3-phase rectifier connection. Six individual cells connected in this manner have a rating of 225 amperes at 65 volts d-c. This is a surprisingly high rating for six of these relatively small germanium cells and is one of the distinct advantages. Such a self-contained power supply can be connected on the d-c output side in series or parallel with other equivalent units to form almost any combination and rating of power supplies required for industry. This voltage rating lends itself to providing the customary 125-volt power supply when two units are connected in series. This arrangement is well within the range of nominal 25-kw rating of packaged power supplies which have been offered for some time, using other metallic disk rectifiers. Fig. 7 gives a vivid comparison of size between a packaged selenium rectifier and the new germanium rectifier, both rated 25 kw, complete with transformers, switches, cooling blowers, and instrumentation. Fig. 8
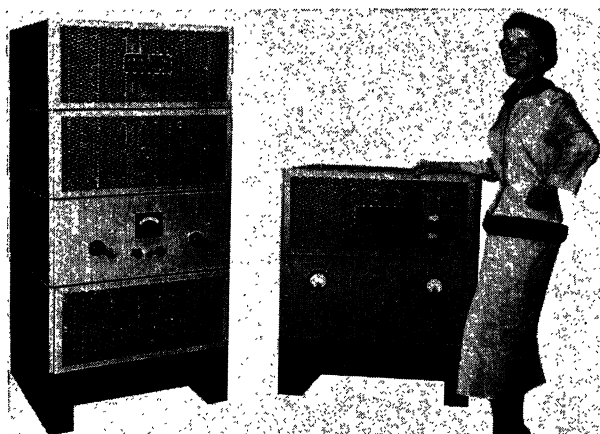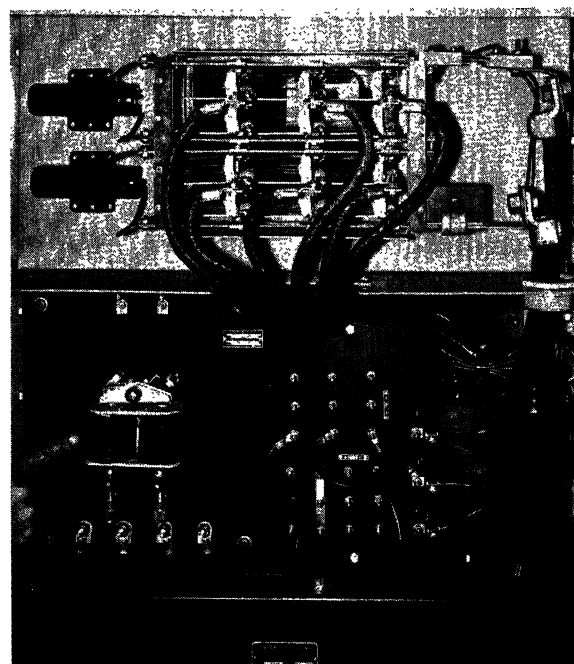
shows the front view of a packaged 25-kw 125-volt germanium power rectifier and Fig. 9 shows a rear view of the same unit, displaying the vertical motor-driven fan which pulls air over the rectifier cells and forces it downward over the transformer, discharging out the bottom of the cubicle.

To summarize, the principal advantages of germanium power rectifiers over other metallic types are: 1. high efficiency; 2. low regulation; 3. good stability; 4. very small reverse current; 5. high inverse voltage rating; and 6. small physical size. Fig. 10 shows a typical efficiency and regulation curve.

Industrial Application

The Metal and Thermit Corporation has become the first purchaser of a germanium power rectifier. This company is engaged in the business of reclaiming tin from tinplate scrap by the electrolytic method, which requires a considerable amount of direct current

at a potential in the order of 115 volts. The load is constant.

Recently this company decided to replace an old motor-generator set. After comparing motor-generator sets, mechanical rectifiers, and other metallic disk rectifiers, a decision was reached to give this new germanium power rectifier a chance to prove the advantages claimed for it. As well as being competitive in first cost, germanium has the advantage of high efficiency, easy installation, quietness of operation and very little maintenance.

A group of four 25-kw packaged power supplies, as described previously, were installed. Fig. 11 shows the installation. These units are connected in parallel to a common d-c bus which will supply a nominal rating of 800 amperes at a voltage up to 125 volts d-c.
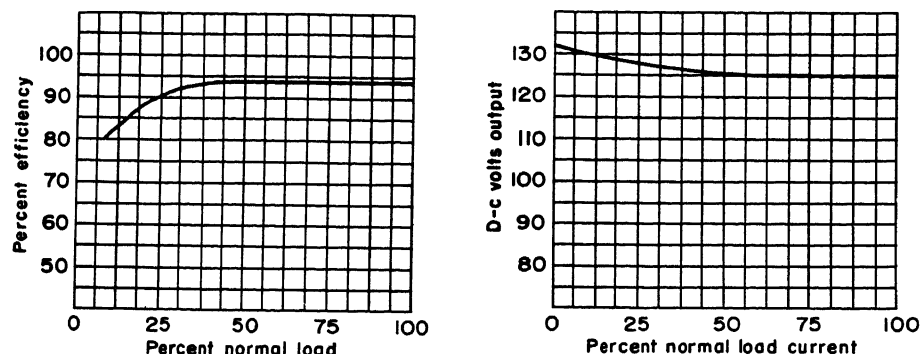
Fig. 10. Efficiency and regulation curve

Character of the Load

The chemical process is beyond the scope of this paper; however, a brief explanation of electrical characteristics will be helpful in understanding the application. In common with most electrolytic processes, the internal resistance of the circuit is quite low. It follows that the initial applied voltage must be quite low to limit inrush current to a reasonable value. A voltage in the order of 25 per cent of final full load voltage is a safe value for use with the rectifier.

As current begins to flow through the electrolyte, a back electromotive force (emf) is produced. This counter voltage rapidly rises to a value equal to the reduced applied voltage. Thus, the inrush current is quickly reduced to practically zero. The applied voltage may now be increased in fairly large steps without causing further excessive inrush current, provided the continuity of the circuit is

maintained. The back emf continues to rise as the applied voltage is increased, finally reaching a maximum value at about 80 per cent of final full load voltage. Beyond this point, it is desirable to have continuous adjustment of applied voltage so it can be set at a value which overcomes load circuit resistance and maintains the desired load current.

Method of Control

When using the d-c generator it had been the practice to turn the field rheostat for minimum voltage (in the order of 35 volts), then close the d-c circuit breaker. From the characteristics of the process previously described, it can be seen this practice would cause a suddent current inrush which would taper off to practically zero as the back emf builds up. With the rectifier, the problem is a little different in that it is desirable to limit inrush current for better

protection of the cells, as well as to minimize tripping of the instantaneous overload. The need for a rather wide range of voltage adjustment is evident, although fortunately rather large steps can be used at the lower values. It was known that the power supply should be about 100 volts for an 800-ampere load current. It was determined by experiment that an initial applied direct voltage of about 25 volts would limit the total inrush current to the order of 200 amperes.

When planning the installation, it was decided that an induction voltage regulator could provide smooth voltage adjustment for a considerable range either side of the normal 100-volt operating point. It did not seem economical to build such a regulator with a range to provide the required starting voltage. At this point, it was discovered that the autotransformer starter from the retired motor-generator set drive motor would be available. By opening the wye of this autotransformer, three series reactors are available and these are inserted in the incoming a-c line adjacent to the voltage regulator to provide a very effective reduced voltage for starting. Fig. 12 shows essentially the schematic connection of the power system as used. The series reactors are not shown here, but they are located just below the contactor.

The voltage regulator is rated 475 volts, 154 amperes, making it suitable for six 25-kw rectifier units. It provides continuous manual control 25 percent
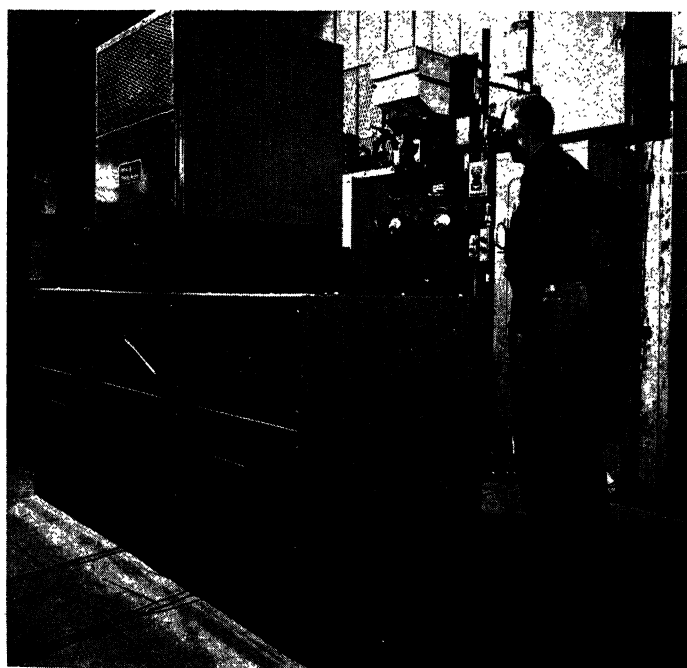
Fig. 11. Complete d-c power supply with germanium rectifiers
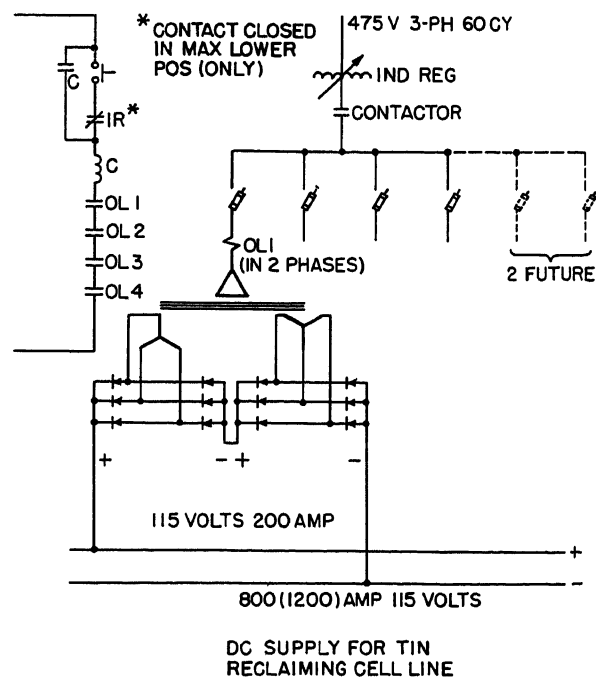
Fig. 12. Diagram of connection of germanium rectifier power system

above and below rated voltage. This range gives 73 volts minimum to 122 volts maximum at the rectifier output terminals.

Proper sequence for start-up is assured by interlocks in the main contactor control circuit. The fans must be running, overloads reset, and voltage regulator at the minimum position before the main contactor can be closed. The manual short-circuiting contactor around the series reactor is always released by its "no voltage release coil" when the main contactor is open.

The first operation is to close the d-c load circuit breaker; then power is applied to line by push-button control which closes the main contactor. This applies the minimum alternating voltage to the rectifiers, resulting in approximately 25 volts d-c at the load. This level of voltage suddenly applied to the load causes a momentary inrush in the order of 200 amperes total current if the

line has been idle for a period long enough for the back emf to disappear. The current quickly drops to practically zero because the back emf soon equals the applied voltage. Short-circuiting of the reactor raises the supply voltage to 73 volts d-c, again causing a current inrush in the order of 200 amperes total. This current rapidly decreases to practically zero as the back emf quickly builds up to a new value. The next step is to increase the d-c bus voltage with the voltage regulator. At about 80 volts on the d-c bus, the load ammeter begins to show steady-state current flow through the load, indicating that the maximum back emf has been reached. The operator continues to increase the d-c bus voltage by means of the voltage regulator until the desired load current is reached. It has been found that 101 volts produces a load current of 800 amperes which is satisfactory operation, leaving ample range in the regulator in either direction to pro-

vide for minor changes in load characteristics.

Operating temperatures of the rectifier cells have been measured at several points under normal load conditions. With an ambient of 32 degrees centigrade, the highest total temperature observed is 49 degrees centigrade.

Refer again to Fig. 11 which shows this installation at work. The four units are arranged in such a manner that two more units can easily be added to bring the total rating to 150 kw, the equivalent of the retired motor-generator set.

A similar germanium rectifier installation has been in operation for several months in a General Electric plant. Several other complete power supply equipments are on order for use in the chemical, aluminum, steel, and titanium industries. Ratings such as 16,000 amperes at 65 volts, 12,000 amperes at 130 volts, and 40,000 at 24 volts are being manufactured for these industries.

Further Effects of the Pole and Zero Locations on the Step Response of Fixed, Linear Systems

ARMEN H. ZEMANIAN
ASSOCIATE MEMBER AIEE

Synopsis: An extension is made to a method of evaluating the effect of the pole and zero locations on the unit step response of fixed, linear systems originally devised by Mulligan.[1] This extension permits a rapid determination of the points at which the step response crosses the final value line for many systems. The first such point is the rise time from zero to the final value.

A SIMPLE and rapid means of obtaining the step response is of great usefulness in evaluating the transient characteristics of a fixed, linear system. Mulligan[1] has devised a very simple though approximate method of obtaining the locations and magnitudes of the maxima and minima of this step response from the pole and zero locations of the corresponding system function without having to determine the total response. The system function is defined as the Laplace transform of the unit step response, for a system which is initially at rest, divided by s where s is the complex frequency argument of the system function. These maxima and minima have abscissas t_i, as shown in Fig. 1. Moreover, this procedure leads to a rapid means of synthesizing a system function when these maxima and minima are prescribed.

The following is an extension of this approach to include those points at which the step response crosses the final value line. These are the points which have abscissas T_i in Fig. 1 and the first such ab-

scissa has a value which is the rise time from zero to the final value (henceforth called "the rise time from zero to one"). The method is applicable only to those system functions which have no poles in the right-half s plane or on the imaginary axis, no zeros at the origin more poles than zeros, and no multiple poles. When a dominant pole approximation (this will be explained later) is used, the last restriction may be lifted so long as the dominant pole pair is not a multiple one. The notation will be the same as that employed by Mulligan.

Notations for Pole and Zero Locations

$-\alpha_k \pm j\beta_k =$ location of Kth pair of complex poles

$-\xi_k \pm j\gamma_k =$ location of Kth pair of complex zeros

$-\rho_k =$ location of Kth real pole

$-\mu_k =$ location of Kth real zero

$\beta_{ok} =$ distance from origin to $-\alpha_k \pm j\beta_k$

$\beta_{ika} =$ distance from $-\alpha_i + j\beta_i$ to $-\alpha_k + j\beta_k$

$\beta_{ikb} =$ distance from $-\alpha_i - j\beta_i$ to $-\alpha_k + j\beta_k$

$\gamma_{ok} =$ distance from origin to $-\xi_k \pm j\gamma_k$

$\gamma_{ika} =$ distance from $-\xi_i + j\gamma_i$ to $-\alpha_k + j\beta_k$

$\gamma_{ikb} =$ distance from $-\xi_i - j\gamma_i$ to $-\alpha_k + j\beta_k$

$\rho_{ik} =$ distance from $-\rho_i$ to $-\alpha_k + j\beta_k$

$\mu_{ik} =$ distance from $-\mu_i$ to $-\alpha_k + j\beta_k$

$\bar{\rho}_{ik} =$ distance from $-\rho_i$ to $-\xi_k + j\gamma_k$

$$\theta_{ik} = \theta_{ika} + \theta_{ikb} = \tan^{-1}\frac{\beta_k - \gamma_i}{\xi_i - \alpha_k} + \tan^{-1}\frac{\beta_k + \gamma_i}{\xi_i - \alpha_k}$$

$$\psi_{ik} = \psi_{ika} + \psi_{ikb} = \tan^{-1}\frac{\beta_k - \beta_i}{\alpha_i - \alpha_k} + \tan^{-1}\frac{\beta_k + \beta_i}{\alpha_i - \alpha_k}$$
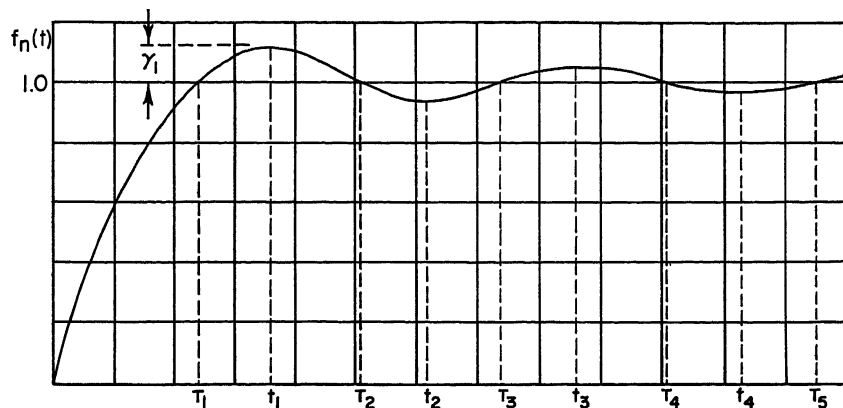
Fig. 1. Illustration of some critical points in the unit step response

$$\psi_k = \tan^{-1}\frac{\beta_k}{\alpha_k}; \; \delta_{ki} = \tan^{-1}\frac{\beta_k}{\mu_i - \alpha_k}$$

$$\phi_{ik} = \tan^{-1}\frac{\beta_k}{\rho_i - \alpha_k}$$

Review of Previously Developed Theory

It is assumed that the system function has n pairs of conjugate complex poles, q real poles, g pairs of conjugate complex zeros, m real zeros, that $2n+q>2g+m$, and that there are only first-order poles and zeros. The notations for the pole and zero locations are illustrated in Fig. 2. The poles and zeros are numbered consecutively according to the magnitude of their real parts starting with the one with the lowest real part. If two or more pairs of conjugate complex poles or zeros have the same real part, then these are numbered consecutively according to the magnitude of their imaginary parts starting with the lowest imaginary part. In this discussion, the quantities designated by Greek letters are either the pole or zero locations in the s plane, distances between these locations, or angles generated by these locations. These symbols are also illustrated in Fig. 2.

The Laplace transform for the unit step response of such a system that is initially at rest is

$$F(s) = K\frac{\prod\limits_{i=1}^{m}(a+\mu_i)\prod\limits_{i=1}^{g}[(s+\xi_i)^2+\gamma_i^2]}{s\prod\limits_{i=1}^{q}(s+\rho_i)\prod\limits_{i=1}^{n}[(s+\alpha_i)^2+\beta_i^2]} \quad (1)$$

The final value of the response is

$$f(\infty) = \lim_{s\to 0} s\, F(s) = K\frac{\prod\limits_{i=1}^{m}\mu_i\prod\limits_{i=1}^{g}\gamma_{oi}^2}{\prod\limits_{i=1}^{q}\rho_i\prod\limits_{i=1}^{n}\beta_{oi}^2} \quad (2)$$

Normalizing the unit step response by this factor, the result is

$$f_n(t) = \frac{f(t)}{f(\infty)}$$

$$f_n(t) = 1 + (-1)^{w+1}\sum_{k=1}^{n}M_k \csc\psi_k\, \epsilon^{-\alpha_k t}\sin$$

$$(\beta_k t+\lambda_k+\psi_k)+(-1)^w\sum_{k=1}^{q}\times$$

$$(-1)^{k+r_k}N_k \epsilon^{-\rho_k t} \quad (3)$$

where

$$M_k = \frac{\prod\limits_{i=1}^{m}\left|\frac{\mu_{ik}}{\mu_i}\right|\prod\limits_{i=1}^{g}\left(\frac{\gamma_{ika}}{\gamma_{oi}}\right)\left(\frac{\gamma_{ikb}}{\gamma_{oi}}\right)}{\prod\limits_{i=1}^{q}\left(\frac{\rho_{ik}}{\rho_i}\right)\prod\limits_{\substack{i=1\\i\neq k}}^{n}\left(\frac{\beta_{kia}}{\beta_{oi}}\right)\left(\frac{\beta_{ikb}}{\beta_{oi}}\right)} \quad (4)$$

$$N_k = \frac{\prod\limits_{i=1}^{m}\left|\frac{\mu_i-\rho_k}{\mu_i}\right|\prod\limits_{i=1}^{g}\left(\frac{\bar{\rho}_{ki}}{\gamma_{oi}}\right)^2}{\prod\limits_{i=1}^{n}\left(\frac{\rho_{ki}}{\beta_{oi}}\right)^2\prod\limits_{\substack{i=1\\i\neq k}}^{q}\left|\frac{\rho_i-\rho_k}{\rho_i}\right|} \quad (5)$$
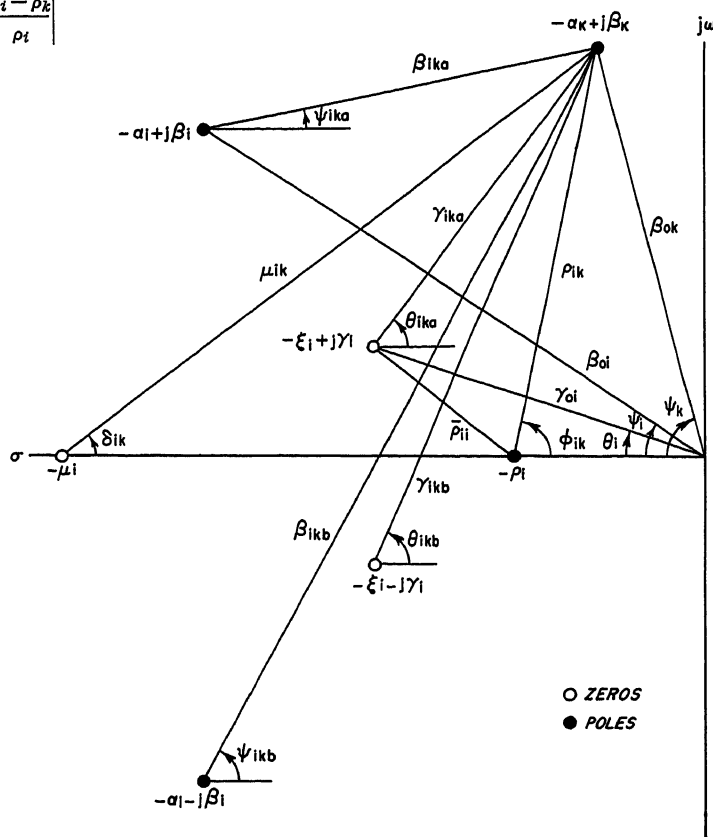
Fig. 2. Illustration of the notations employed for the pole and zero locations

$$\lambda_k = \sum_{i=1}^{g}\theta_{ik} - \sum_{\substack{i=1\\i\neq k}}^{n}\psi_{ik} + \sum_{i=1}^{m}\delta_{ik} - \sum_{i=1}^{q}\phi_{ik} \quad (6)$$

and where w and r_k are the number of real zeros in the right-half plane and the number of real zeros to the right of ρ_k respectively.

Some simple graphical means exist for the evaluation of the factors M_k, N_k, and λ_k. Since in the forthcoming extension of the theory to include the points $[T_i, f(\infty)]$. it is only necessary to evaluate λ_k, the graphical methods applying to M_k and N_k will not be reviewed.

For every complex pole pair, there is an angle λ_k, which is the sum of the angles due to all other poles and zeros as shown in Fig. 2. If the pole and zero locations are normalized by dividing every coordinate by β_k and then entered into Fig. 3, the contributions to λ_k of any pole or zero may be obtained. This chart may be constructed from the consideration that, if the pole $(-\alpha_i+j\beta_i)$ moves along a circle of radius $\beta_k \csc\psi_{ik}$ with center at $\sigma = -(\alpha_k+\beta_k\cot\psi_{ik})$, then the pair of poles $(-\alpha_i\pm j\beta_i)$ contributes a constant angle ψ_{ik} to λ_k. Using expression 6 to evaluate λ_k, elements to the right of the abscissa of $(-\alpha_k\pm j\beta_k)$ produce a negative contribution to ψ_{ik} or θ_{ik} and those to the left produce a positive contribution. Moreover, real elements to the left of $-\alpha_k$ contribute angles of one-half the value

indicated while those to the right of $-\alpha_k$ contribute π minus half the values indicated. Finally the contributed angle of an n order element is n times the angle of a first order element.

This is the point of departure from which this approach will be extended to include the points wherein $f_n(T_i) = f_n(\infty)$. The assumption is made that such points exist. The theorem proposed by Mulligan, which follows, may be useful as an indication of their existence.

Theorem: The transient response to a step function input of a linear, fixed lumped and stable system having no real poles between the origin and the real part of the first pair of complex poles cannot be monotonic.

Furthermore, a theorem proven in reference 2 may similarly be of use if sufficient information about the frequency response of the system is available.

Theorem: If the magnitude or real part of a low-pass system function of a linear, fixed, lumped and stable system vanishes with increasing frequency and has a value at any frequency greater than its value at zero frequency (i.e., if it "peaks"), then the corresponding step response cannot be monotonic.

Development of New Theory

Setting $f_n(T_i) = f_n(\infty) = 1$ in equation 3, the following expression is obtained

$$0 = (-1)^{w+1} M_1 \csc \psi_1 \epsilon^{\alpha_1 T_i} \sin$$
$$(\beta_1 T_1 + \lambda_1 + \psi_1) + (-1)^{w+1} \times$$
$$\sum_{k=2}^{n} M_k \csc \psi_k \epsilon^{-\alpha_k T_1} \sin$$
$$(\beta_k T_i + \lambda_k + \psi_k) +$$
$$(-1)^{w} \sum_{k=1}^{q} (-1)^{k+r_k} N_k \epsilon^{-\rho_k T_i} \quad (7)$$

wherein the first term is written separately from the first summation. Dividing through by $(-1)^{w+1} M_1 \csc \psi_1 \, \epsilon^{-\alpha_1 T_i}$

$$0 = \sin (\beta_1 T_i + \lambda_1 + \psi_1) +$$
$$\sum_{k=2}^{n} \frac{M_k \csc \psi_k}{M_1 \csc \psi_1} \epsilon^{-\alpha_k T_i} \sin (\beta_k T_i + \lambda_k + \psi_k) +$$
$$\sum_{k=1}^{q} (-1)^{k+r_k+1} \frac{N_k}{M_1} \frac{1}{\csc \psi_1} \times \epsilon^{-\rho_k T_i} \quad (8)$$

where $\alpha_k^1 = \alpha_k - \alpha_1$ and $\rho_k^1 = \rho_k - \alpha_1$.

Now the quantity X_i will be defined by the following expression

$$\beta_1 T_i = c\pi - \lambda_1 - \psi_1 + x_i \quad (9)$$

where $c = d + i$ and d is the largest integer (positive, negative, or zero) for which $(d\pi - \lambda_1 - \psi_1)$ is negative. It is the value by which the $\beta_1 T_i$ for the actual response differs from the $\beta_1 T_i$ resulting from the approximation of the unit step response

by the first and second terms of expression 3. Normalizing the s plane pole and zero locations by letting $\beta_1 = 1$ and inserting this value of T_1 into expression 8, the following is obtained

$$\sin x_i = (-1)^{c+1} \sum_{k=2}^{n} \frac{M_k \csc \psi_k}{M_1 \csc \psi_1} \times$$
$$\epsilon^{-\alpha_k^1(c\pi x - \lambda_1 - \psi_1 + x_i)} \sin [\beta_k(c\pi - \lambda_1 -$$
$$\psi_1 + x_i) + \lambda_k + \psi_k] + (-1)^c \sum_{k=1}^{q} \times$$
$$(-1)^{k+r_k} \frac{N_k}{M_1 \csc \psi_1} \epsilon^{-\rho_k^1(c\pi - \lambda_1 - \psi_1 + x_i)}$$
$$(10)$$

This expression provides a means of calculating the quantity x_i by a trial and error process once the pole and zero locations of a particular system are known. The procedure is to assume the x_i on the right-hand side of expression 10 as zero and then calculate the $\sin x_i$ by the resulting expression. Inserting the new value of x_i into this expression, the process may be repeated. This is continued until the x_i do not vary. Then expression 9 may be used to calculate the T_i. The first point T_1, which is the value of rise time from zero to one for the response, is obtained when c is set equal to $d+1$.

Under certain conditions on the pole and zero locations, the x_i will be comparatively small and the T_i may be calculated without the necessity of determining the x_i. That is, the unit step response is approximated by the first two terms on the right-hand side of equation 3. The second term corresponds to the dominant pair of poles (that complex pair which is closest to the $j\omega$ axis in the s plane and closest to the σ axis if there is more than one such complex pole pair). These conditions will now be determined. Expression 10 may be rewritten as follows

$$\sin x_i = (-1)^{c+1} \sum_{k=2}^{n} \frac{1}{K_{Mki}} \times$$
$$\sin [\beta_k(c\pi - \lambda_1 - \psi_1 + x_i) + \lambda_k + \psi_k] +$$
$$(-1)^c \sum_{k=1}^{q} (-1)^{k+r_k} \frac{1}{K_{Nki}} \quad (11)$$

where

$$K_{Mki} = \frac{M_1 \csc \psi_1}{M_k \csc \psi_k} \epsilon^{\alpha_k^1(c\pi - \lambda_1 - \psi_1 + x_i)}$$

$$K_{Nki} = \frac{M_1 \csc \psi_1}{N_k} \epsilon^{\rho_k^1(c\pi - \lambda_1 - \psi_1 + x_i)}$$

If the K_{Mki} and the K_{Nki} are large, the contributions to $\sin x_i$ will be small and a good approximation will result. Therefore, speaking qualitatively, the conditions for a good approximation are as follows:

1. The greater α_k^1 and ρ_k^1 are (i.e., the further to the left in the s plane the other

poles are as compared to the dominant pair) the better will be the approximation.

2. The greater the ratio of $\sin \psi_k / \sin \psi_1$ (i.e., the lower in the s plane the first complex pole pair is and the higher the other complex poles are) the better will be the approximation.

3. The larger the ratios M_1/M_k and M_1/N_k are, the better will be the approximation. From expressions 4 and 5 it can be seen that these ratios may be written as follows

$$\frac{M_1}{M_k} = \frac{\prod\limits_{i=1}^{m} \mu_{i1} \prod\limits_{i=1}^{g} \left(\frac{\gamma_{i1a}}{\gamma_{ika}}\right)\left(\frac{\gamma_{i1b}}{\gamma_{ikb}}\right)}{\prod\limits_{i=1}^{q} \rho_{i1} \prod\limits_{\substack{i=2 \\ i\neq k}}^{n} \left(\frac{\beta_{i1a}}{\beta_{ika}}\right)\left(\frac{\beta_{i1b}}{\beta_{ikb}}\right)}$$

$$\frac{M_1}{N_k} = \frac{\prod\limits_{i=1}^{m} \left|\frac{\mu_{i1}}{\mu_i - \rho_k}\right| \prod\limits_{i=1}^{g} \left(\frac{\gamma_{i1a}}{\bar{\rho}_{kt}}\right)\left(\frac{\gamma_{i1b}}{\bar{\rho}_{kt}}\right)}{\prod\limits_{i=2}^{n} \left(\frac{\beta_{i1a}}{\rho_{kt}}\right)\left(\frac{\beta_{i1b}}{\rho_{kt}}\right) \prod\limits_{\substack{i=1 \\ i\neq k}}^{q} \left|\frac{\rho_{i1}}{\rho_i - \rho_k}\right|}$$

For these ratios to be large the factors for the zeros in the numerator should be large and the factors for the poles in the denominator shall be small, speaking qualitatively. The following conditions would favor such ratios:

3a. There should be no zeros too close to the dominant poles (large μ_{i1}, γ_{i1a}, and γ_{i1b}).

3b. Bunching of the zeros would be preferable (small μ_{ik}, γ_{ika}, γ_{ikb}, $|\mu_i - \rho_k|$, and $\bar{\rho}_{kt}$).

3c. The poles should be well separated (large ρ_{ik}, β_{ika}, β_{ikb}, and $|\rho_i - \rho_k|$).

As was to be expected, the conditions which lead to a good approximation of the actual T_i by a dominant pole approximation are the same as those which lead to a good approximation of the maxima and minima. The effect of the various factors of K_{Mki} and K_{Nki} have been determined quantitatively in Mulligen's paper and will not be repeated here. It is to be noted that large separation between the poles, especially horizontal separation between the first pair of poles and the others, with no complex poles too near the real axis other than the dominant pair and with no zeros too close to the first pair of poles are conditions which favor an approximation of the unit step response by the first two terms in equation 3. Such an approximation leads to considerable simplification in the analysis of complex systems.

As an example of the ease with which these points T_i may be obtained, the Doba network will be considered. This system has a driving point impedance having a dominant pair of complex poles at $(-1.625 \pm j1.740)$, a real pole at (-2.34) and a pair of complex zeros at $(-2.30 \pm j2.83)$. The corresponding unit
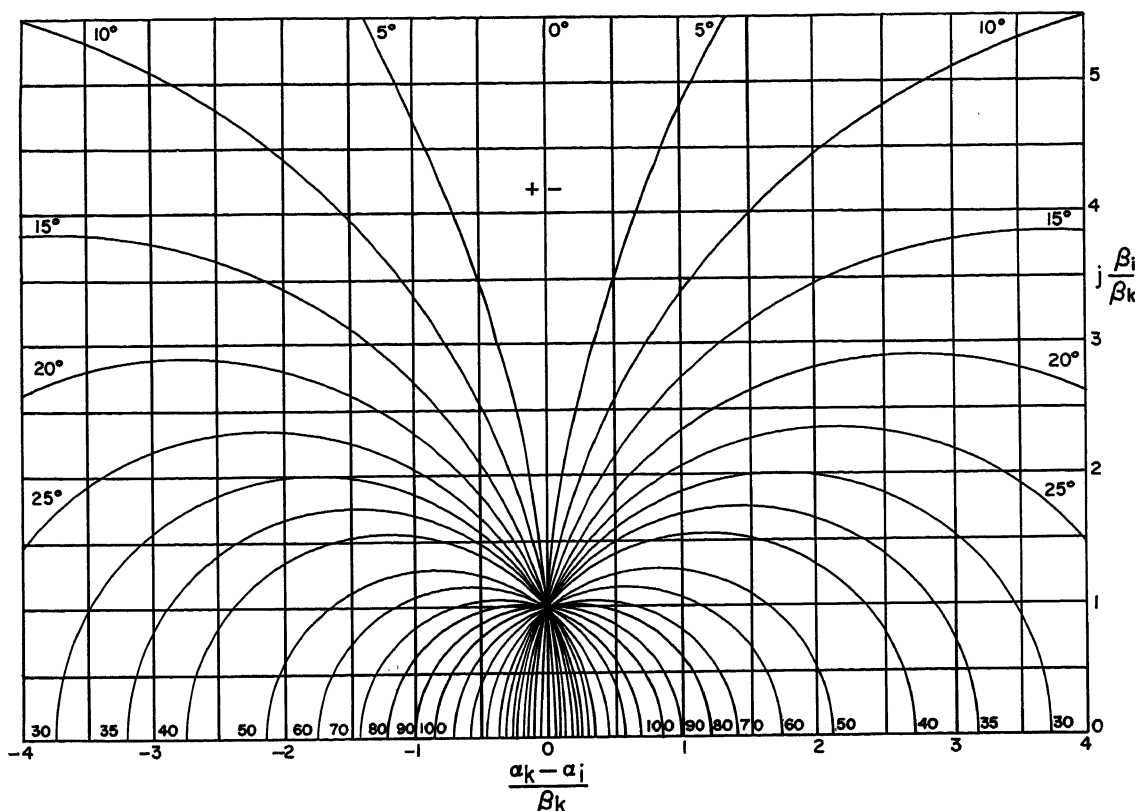
step response is

$$f_n(t) = 1 - 0.760\epsilon^{-1.625t}\sin$$
$$(1.740t + 0.046) - 0.966\epsilon^{-2.34t}$$

The angle λ_1 may be obtained either by calculating its components according to equation 6 or by normalizing the pole and zero locations and using Fig. 3. Its value is -0.773 radian and the value for ψ_1 is 0.819 radian. Use of equation 9 where x_i is arbitrarily taken as zero then yields the approximate value for the first point T_1 as 1.78 seconds. Its actual value is 1.99 seconds. The approximate value for the second point is 3.59 seconds while the actual value is 3.55 seconds. The approximations for the successive points are even better.

Finally the rise time from zero to one and the first overshoot may be exactly determined once the pole and zero locations of a system function and the final value of the corresponding unit step response are known for a one complex pole pair and one real zero system. A chart yielding these quantities and finding use either in the analysis or synthesis of such systems may be found in reference 3 wherein the foregoing material is also treated in greater detail.

Conclusions

The procedure described in this paper permits the determination of the points where the step response crosses its final value line without having to determine the entire response. This method involves a trial and error process. However, if the pole and zero locations favor a dominant pole approximation, these points can be approximately determined by use of equation 9 setting $x_i = 0$. The quantities β_1, λ_1, and ψ_1 are determined entirely by the pole and zero locations.

References

1. THE EFFECT OF POLE AND ZERO LOCATIONS ON THE TRANSIENT RESPONSE OF LINEAR DYNAMIC SYSTEMS, J. H. Mulligan, Jr. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 37, May 1949, pp. 516–29.

2. BOUNDS EXISTING ON THE TIME AND FREQUENCY RESPONSES OF VARIOUS TYPES OF NETWORKS, A. H. Zemanian. *Ibid.*, vol. 42, May 1954, pp. 835–39.

3. INVESTIGATION OF THE TRANSIENT RESPONSE OF LINEAR SYSTEMS, A. H. Zemanian. *Doctoral Dissertation*, New York University, New York, N. Y. 1953. chap. 6.

An Eddy-Current Braking Crane-Hoist Controller with Variable Brake Excitation

H. J. RATHBUN
ASSOCIATE MEMBER AIEE

Synopsis: An a-c crane-hoist controller for a wound-rotor induction motor and an eddy-current brake is described. The brake serves as an artificial load for the motor. Emphasis is placed on the desirability of using speed-responsive excitation for the brake to cause its torque to increase so rapidly with speed that flat speed-load characteristics are obtained. A magnetic amplifier responsive to a signal voltage taken from the secondary circuit of the motor is a simple and reliable means for controlling the brake excitation.

THE polyphase wound-rotor induction motor is used extensively for crane hoists. Many of these hoist drives, particularly those handling loads which must be accurately positioned, should be operable at several different subsynchronous speeds which are substantially independent of the load on the hook.

With resistance added to its secondary circuit, a polyphase wound-rotor induction motor, when operating in the direction of its torque with balanced voltages on its primary, runs at speeds materially below its synchronous speed only if it is driving a substantial load. The actual load on a crane-hoist motor, however, varies throughout a wide range. It can be a small positive load as when hoisting or lowering an empty hook, a large positive load as when hoisting a heavy load, a small negative or overhauling load as when lowering a light load, or a large overhauling load as when lowering a heavy load. Consequently, it has been found desirable for certain a-c crane-hoist applications to use an artificial load which so loads the motor that its speed may be changed materially by changing the resistance of its secondary circuit regardless of the actual load on the hook. For example, mechanical load brakes have been used on cranes to provide an

artificial load for a wound-rotor motor permitting heavy hook loads to be lowered slowly, and electric load brakes have been coupled directly to the motor shaft to load the motor artificially during hoisting as well as lowering. The electric load brake can be a generator having an external resistive circuit or it can be a generator having no external circuit such as an eddy-current brake.[1,2] The eddy-current brake has advantages in simplicity and economy. A controller for such a combination of a motor and an eddy-current brake is called an eddy-current braking hoist controller.

Motor Operation

In so far as the motor is concerned, an eddy-current braking hoist controller is usually a plain-reversing magnetic controller operated by a multiposition reversing master switch. If it is necessary to lower very heavy loads at a creeping speed, the torque of the eddy-current brake can be augmented by unbalanced voltage braking torque in the first lowering speed point. A typical plain-reversing eddy-current braking hoist controller providing five hoisting and five lowering speed points controls the secondary circuit of the motor to provide the speed-torque curves of Fig. 1. Torque is plotted in per cent of the rated full-load torque of the motor and speed is plotted in per cent of synchronous speed. Fig. 2 is an elementary wiring diagram of the external secondary circuit.

In the first lowering speed point of the master switch, all of the acceleration contactors are open, and the motor operates along speed-torque curve 1 in the power-lowering quadrant of Fig. 1. The first acceleration contactor *1A* closes in the second lowering speed point, changing the motor performance to that indicated by lowering curve 2. In the third lowering speed point, the second acceleration contactor *2A* closes, and the operation is in accordance with lowering curve 3. In the fourth lowering speed point, the third acceleration contactor *3A* closes, and the

operation is in accordance with lowering curve 4. After the master switch reaches the fifth lowering speed point, the fourth and the last acceleration contactors *4A* and *5A* close in sequence. Closure of the acceleration contactor *4A* causes the motor to operate along an intermediate acceleration curve indicated by a broken line in Fig. 1. Only contactors *4A* and *5A* need to be controlled by acceleration relays while lowering.

In the first hoisting speed point, contactor *1A* closes, and the motor operates along speed-torque curve 1 in the power hoisting quadrant. Contactor *2A* closes in the second hoisting speed point, but no contactors operate upon movement of the master switch to the third hoisting speed point. The hoisting speed-torque curve 2–3 is for operation in these two speed points. In the fourth hoisting speed point, contactor *3A* closes, and in the fifth hoisting speed point contactors *4A* and *5A* close in sequence, closure of contactor *4A* causing the motor to operate along an intermediate acceleration curve indicated by a broken line. Contactors *3A*, *4A*, and *5A* should be controlled by acceleration relays while hoisting. When the motor is controlled as just described, the eddy-current brake is energized in the first two hoisting speed points and in the first four lowering speed points and is de-energized in all other speed points.

Brake Construction and Operation

A diametric cross section of a fan-cooled eddy-current brake suitable for crane-hoist service and arranged for mounting on a motor frame or other vertical surface is shown in Fig. 3. A toroidal coil causes interdigitated stator teeth to be of alternate polarity. The rotor rim cuts the flux between these teeth causing eddy currents to flow in the rim. The retarding torque of the brake is of course zero at standstill and increases with speed and with excitation.[3] Similar brakes are also available for floor mounting. If the brake is operated with different values of constant excitation, its speed-torque performance is as shown in Fig. 4. When the excitation is constant, the torque of the brake increases rapidly with speed up to about 500 rpm, but at higher speeds increases very slightly with speed.

The crane-rated excitation of the brake is well below the saturation value, and consequently the brake torque at any given speed is approximately proportional to the excitation. For most crane-hoist applications, the crane-rated excitation of the brake provides a torque at the
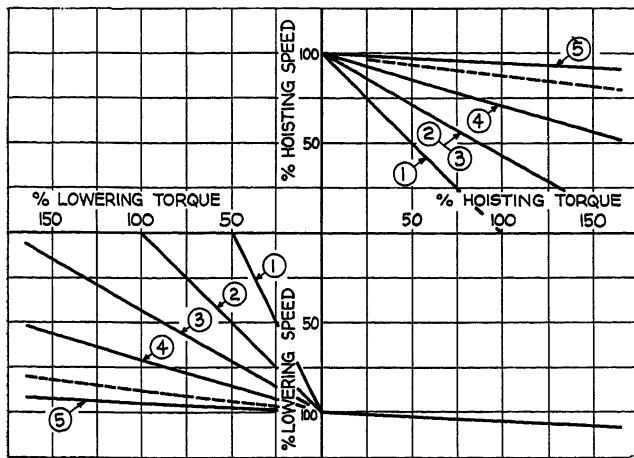
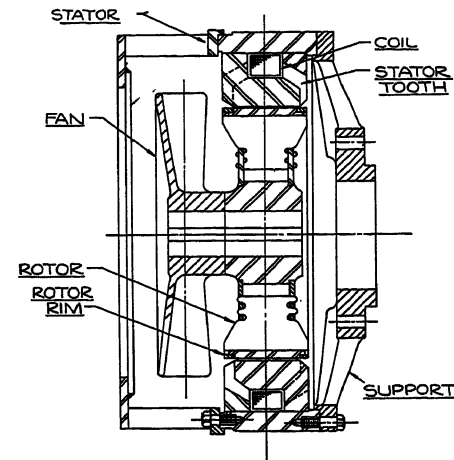Fig. 1 (left). Speed-torque curves of the motor alone

Fig. 3 (right). Diametric cross section of an eddy-current crane brake

synchronous speed of the motor which is greater than the rated full-load torque of the motor, but, for purposes of explanation, it will be assumed in all cases that the crane-rated excitation of the brake is such that its torque at synchronous speed is equal to the rated full-load torque of the motor. A synchronous speed of 1,200 rpm is assumed. Consequently, base torque in pounds-feet is the same for both motor and brake and the base excitation is the excitation producing a brake torque at 1,200 rpm equal to the rated full-load torque of the motor.

Operation of Motor and Brake Combined

Since the motor and the eddy-current brake are directly coupled together, their torques are directly opposed whenever the brake is excited and the motor is operating as a motor in the direction of its torque. A speed-torque curve of the motor alone for an exemplary lowering speed point and a speed-torque curve of the brake alone at two different values of constant excitation are shown in Fig. 5. At a certain selected lowering speed, the motor exerts a lowering torque T_M and the brake operating with base excitation exerts a retarding or hoisting torque T_B. Consequently, the resultant torque at the common motor and brake shaft is the algebraic sum of torques T_M and T_B. Since, for the speed selected, T_B is greater than T_M, the resultant torque T_R is a hoisting torque. After acceleration or deceleration, an overhauling load exerting a per-unit lowering torque of T_R on the

common shaft therefore lowers at the selected speed. By adding the motor torque to the brake torque algebraically for other selected speeds, the speed-torque performance at the common motor and brake shaft can be predicted for those speed points in which the brake is excited.

Fig. 5 shows two resultant speed-torque curves. The motor in each instance is operating with the same secondary resistance. The resultant curve shown by a solid line is for performance at the common motor and brake shaft when the brake excitation is constant at its base value, and the resultant curve shown by a broken line is for operation when the per-unit excitation of the same brake is increased to 150 per cent. A comparison of these two resultant curves shows that increasing the brake excitation from one constant value to another constant value merely decreases the speed for all loads and does not improve the speed regulation. A similar result occurs when a larger brake is substituted.

If the excitation of the eddy-current brake is increased continuously with speed, its speed-torque curve for an exemplary lowering speed point becomes as shown in Fig. 6. Combining the speed-

torque curve of the brake, when its excitation is varied, with the speed-torque curve of the motor gives the resultant speed-torque curve shown by the broken line. The resultant speed-torque curve shown by the solid line is obtained by combining the speed-torque curve of the motor with the speed-torque curve of the same brake operating with constant excitation at base value. A comparison of these two resultant curves shows that improved speed regulation is obtained by varying the brake excitation.

Speed-Responsive Signal Voltage

In order to increase the brake excitation with speed, a speed-responsive signal must be obtained. Conveniently, this signal can be a voltage taken from the external secondary circuit of the motor. Referring again to Fig. 2, if E_s is the voltage across the slip rings, the phase voltage E_p across each of the legs of the balanced wye-connected secondary resistor is

$$E_p = \frac{E_s}{\sqrt{3}} \text{ volts} \qquad (1)$$

An advantageous signal voltage is the voltage drop between the resistor taps $R4$
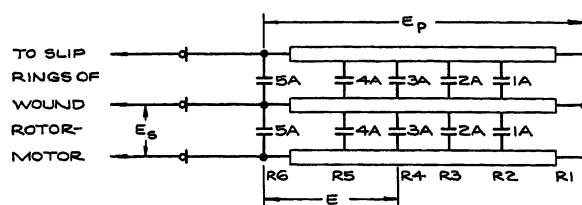
Fig. 2 (below). Elementary wiring diagram of the external secondary circuit of the motors

Fig. 4 (right) Speed - torque curves of an eddy - current crane brake with different values of constant excitation
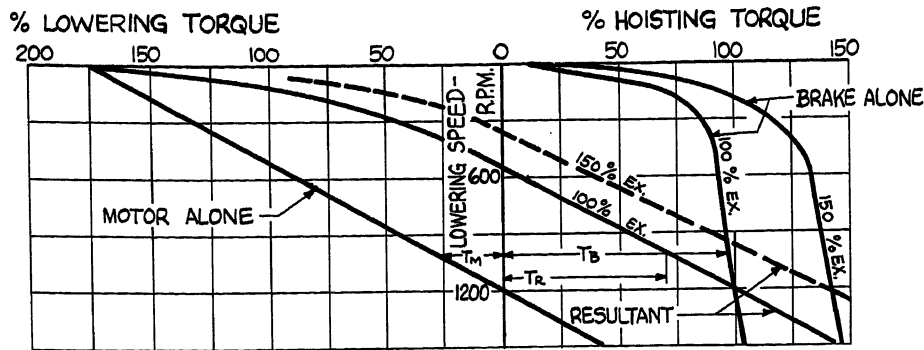
% LOWERING TORQUE   **% HOISTING TORQUE**
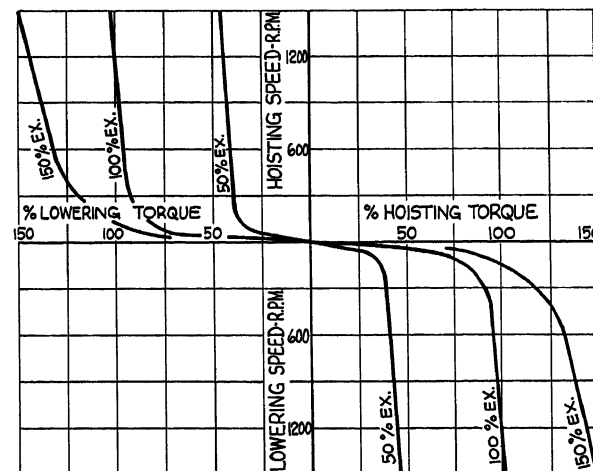
Fig. 5. Effect of increasing brake excitation from one constant value to a second constant value

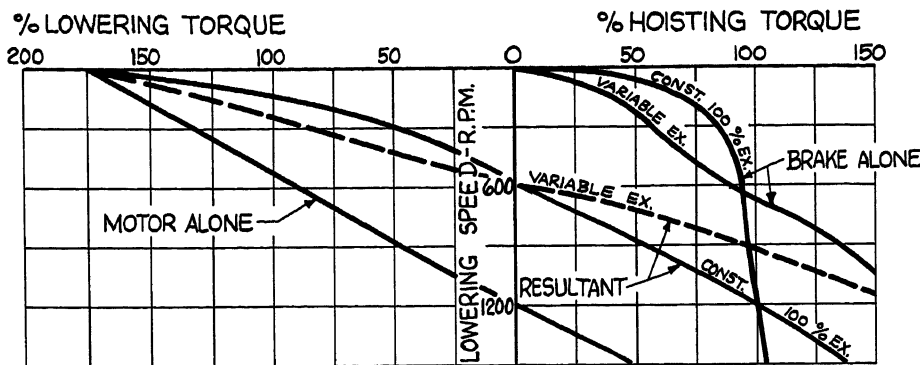**% LOWERING TORQUE**   **% HOISTING TORQUE**

Fig. 6. Comparison between operation with variable and constant brake excitation

and *R6* of one leg of the resistor. This signal voltage *E* is a proportionate part of the phase voltage E_p and depends upon the ratio of the resistance of *R4–R6* to the total effective resistance of the same leg of the secondary resistor. Since the voltages E_s and E_p have a fixed relation, the ratio of the voltage *E* to the voltage E_s also depends upon the ratio of the resistance of *R4–R6* to the total effective resistance of the leg as expressed in the

equation

$$E = \frac{E_s}{\sqrt{3}} \times \frac{\text{resistance of } R4\text{-}R6}{\text{total effective resistance of the leg}}$$
$$\text{volts} \quad (2)$$

The total effective resistance of the leg changes upon operation of the acceleration contactors *1A*, *2A*, and *3A* whereas the resistance from *R4* to *R6* does not.
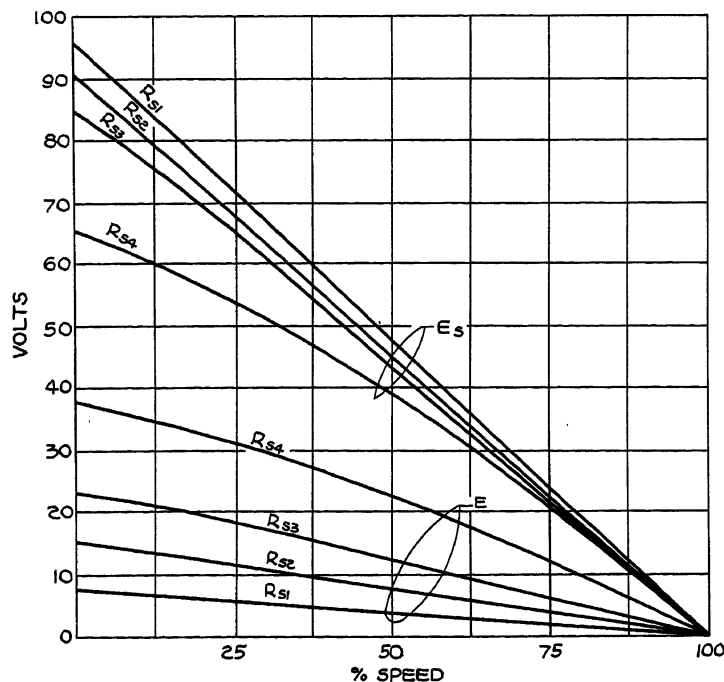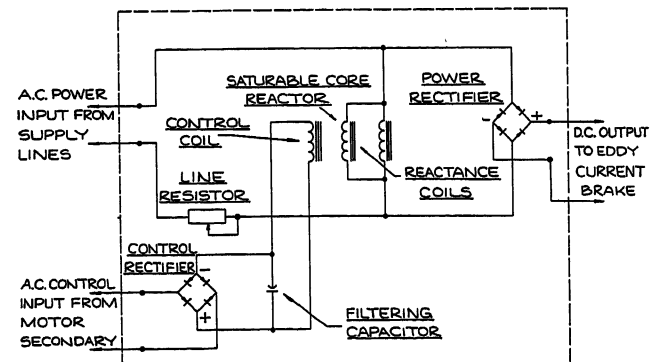
Consequently, as the acceleration contactors *1A*, *2A*, and *3A* close, the ratio between the signal voltage *E* and the ring voltage E_s becomes greater.

The ring voltage E_s decreases both in magnitude and frequency as the motor speed increases. It becomes zero at synchronous speed. Because of the increased internal voltage drop caused by the increased secondary current that flows whenever the secondary resistance is reduced, the magnitude of the ring voltage E_s also decreases at any given subsynchronous speed upon closure of an acceleration contactor.

The upper four curves of Fig. 7 show the variation of the ring voltage E_s with speed for four different values of secondary resistance R_s that are used in the first four lowering speed points respectively. The intermediate two of these resistance values are also used in the first two hoisting speed points respectively. By substituting values of E_s from the curves into equation 2, the variation of the signal voltage *E* with speed is obtained. The lower four curves of Fig. 7 show how the signal voltage *E* varies with speed for the same four resistance values.

The signal voltage *E* gives an indication of motor speed but it decreases with motor speed, whereas the excitation of the eddy-current brake in any selected speed point should increase with speed. The signal voltage also increases as resistance is removed from the secondary circuit whereas for the desired crane performance the range of excitation of the brake should become less as the master switch is moved to the faster speed points. To accomplish the desired inversions in the relations between the signal voltage and speed and between the signal voltage and the changing of the secondary resistance, a magnetic amplifier has been developed. This amplifier provides an output voltage

Fig. 7 (left). Variation of ring voltage E_s and signal voltage E with speed for different values of secondary resistance

Fig. 8 (below). Elementary wiring diagram of the magnetic amplifier

58   *Rathbun—Eddy-Current Braking Crane-Hoist Controller*   MARCH 1955

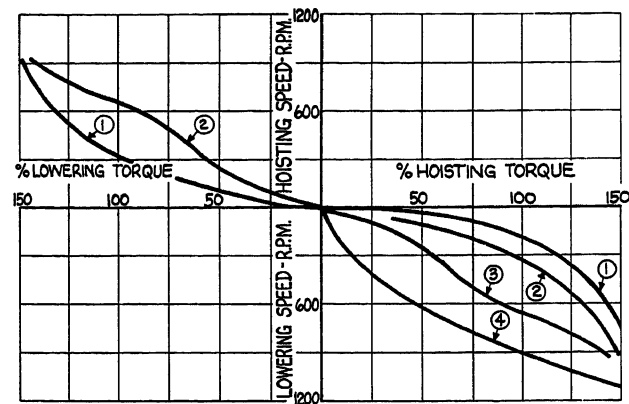Fig. 9. Load characteristic of the magnetic amplifier

Fig. 10. Speed-torque curves of an eddy-current brake alone when its excitation is varied

Fig. 11. Speed-torque relations of the motor and brake combined
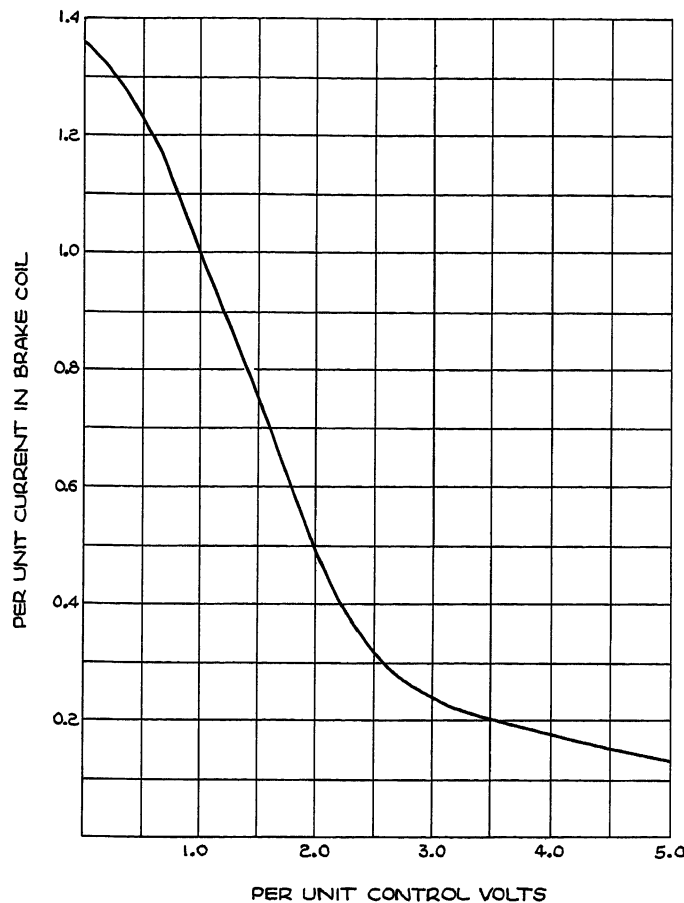
that varies inversely with respect to its control voltage.

Magnetic Amplifier

Fig. 8 is an elementary wiring diagram of the magnetic amplifier which forms a part of the complete eddy-current braking hoist controller. A saturable core reactor in the amplifier has a single coil and two reactance coils. The reactance coils are in parallel with each other and both are connected in series with a line resistor across the power input terminals of the amplifier. The d-c output voltage of a power rectifier connected in parallel with the reactance coils is applied to the coil of the eddy-current brake.

A transformer or potentiometer can be used to change the signal voltage E to a control voltage of suitable value for the amplifier. The control voltage is impressed on the control input terminals of the amplifier and is rectified by the control rectifier. The d-c output of the control rectifier is supplied to the control winding of the saturable reactor.

Fig. 9 is the load characteristic of the magnetic amplifier and shows the relation between the control voltage and the output current when the coil of an eddy-

current brake is connected across the output terminals. The control voltage and the brake current are in a per-unit system in which the base voltage is the voltage which causes base brake excitation, and the base brake excitation is the excitation which causes the brake to exert its base torque at 1,200 rpm. When the control voltage is large, the core of the saturable reactor is saturated, the reactance coils have low impedance, the voltage drop across the line impedance is large, and the output voltage and current are a minimum. Upon a decrease in the control voltage, the impedance of the reactance windings increases, the voltage drop across the line impedance decreases, and the output voltage and current increase. When there is no current in the control coil, the impedance of the reactance coils is a maximum and the output voltage is a maximum.

When control voltages which vary like the signal voltages E of the lower group of the curves of Fig. 7 are impressed on the amplifier, the brake excitation changes with speed and with operation of the acceleration contactors to provide the retarding speed-torque characteristics of
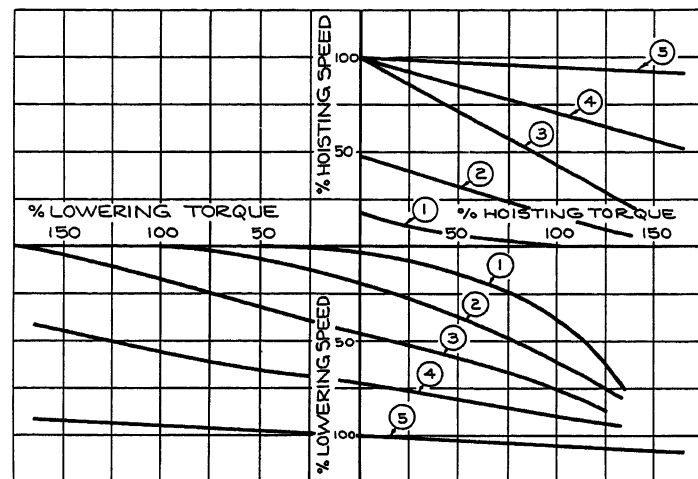
Fig. 10. Lowering curves 1 through 4 show how the brake torque changes in the first four lowering speed points respectively, and hoisting curves 1 and 2 show how the brake torque varies in the first two hoisting speed points respectively. The speed-torque relations at the common motor and brake shaft can be predicted by combining the appropriate curves of Fig. 1 and Fig. 10 in the manner explained in connection with Fig. 5. The results are shown in Fig. 11.

Description of Complete Controller

Fig. 12 is an elementary wiring diagram of the eddy-current braking hoist controller which controls the motor and brake to give the performance of Fig. 11. Frequency relays are used to control acceleration.[4,5] A resistor is used as the line impedance of the magnetic amplifier, and in the "off" position of the master switch this resistor has an effective value which limits the brake current to an amount just sufficient to insure closing of the a-c/d-c relay. The a-c/d-c relay provides an interlock between the eddy-current brake circuit and the controller so that over-

Fig. 13. Performance of a crane-hoist when controlled by the eddy-current braking hoist controller

Fig. 12. Elementary wiring diagram of the eddy-current braking hoist controller
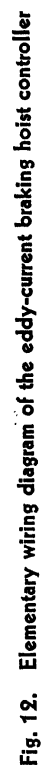
speeding cannot occur as a result of loss of current in the brake circuit. Closure of the a-c/d-c relay completes the circuit to the coil of the undervoltage relay UV.

When the master switch is moved to the first hoisting speed point, relay $1BR$ closes to release the friction brake and contactors H and M close to cause the motor to exert hoisting torque. The acceleration contactor $1A$ also closes instantly. Contacts on the master switch operate to eliminate a portion of the line resistor causing it to have its normal operating value. In the second hoisting speed point, the acceleration contactor $2A$ closes instantly to increase the motor torque and decrease the brake torque. The time delay relay TR also becomes energized in this speed point and closes its contacts instantly to by-pass the contacts of the a-c/d-c relay. In the third hoisting speed point, power is removed from the eddy-current brake at the master switch and the a-c/d-c relay opens. After the acceleration relay $3AR$ operates with the master switch in the fourth hoisting speed point, the acceleration contactor $3A$ closes. The acceleration relays $4AR$ and $5AR$ control the closure of contactors $4A$ and $5A$ after the master switch is moved to the last hoisting speed point. The contacts of the relay TR are delayed in opening so that, upon return of the master

switch to the second or first hoisting speed point, relay UV is held closed through the contacts of relay TR until the a-c/d-c relay is again energized.

In lowering, contactors L and M close to cause the motor to exert lowering torque. Contactor $1A$ remains open in the first speed point and closes instantly in the second speed point. Contactor $2A$ closes instantly in the third speed point and contactor $3A$ closes instantly in the fourth speed point. This causes the motor torque to increase in steps and the brake torque to decrease in steps. When the last speed point is reached, power is removed from the eddy-current brake at the master switch and contactors $4A$ and $5A$ close in sequence under control of relays $4AR$ and $5AR$. If the eddy-current brake should become de-energized inadvertently while the master switch is in the first two lowering speed points, the consequent opening of the a-c/d-c relay causes opening of the undervoltage relay UV which removes power from the motor and causes the friction brake to set. Relay TR is energized in the third lowering speed point and closes its contacts to by-pass the contacts of the a-c/d-c relay. Should the brake excitation fail while the master switch is in the third or fourth lowering speed points, the secondary resistance is low enough to prevent danger-

ous overspeeding. The circuit through the normally closed contacts of the a-c/d-c relay to the operating coils of contactors $4A$ and $5A$ insures that these contactors remain closed until the eddy-current brake is again energized after movement of the master switch from the fifth lowering speed point.

Fig. 13 shows the results of a test made on a 10-ton crane equipped with a controller like that of Fig. 12. The performance of the crane for both hoisting and lowering is given with the load on the crane hook in tons plotted against the hook speed in feet per minute.

References

1. UNDERSYNCHRONOUS SPEED- CONTROL OF POLYPHASE ASYNCHRONOUS MOTORS WITH D-C BRAKING GENERATORS FOR CRANE HOIST DRIVES, Von R. Aschenbrenner. *Elektrotechnik und Maschinenbau*, Vienna, Austria, vol. 58, no. 11–12, Mar. 9, 1940, pp. 115–19.

2. MAGNETIC TORQUE A-C CRANE CONTROL, F. W. Wendelburg, F. M. Blum. *Proceedings*, Association of Iron and Steel Engineers, Pittsburgh, Pa., 1949, pp. 435–44.

3. ELECTRIC EDDY CURRENT MACHINERY, J. B. Winther. *Electrical Engineering*, vol. 66, July 1947, pp. 643–46.

4. TWO NEW METHODS OF ACCELERATING ELECTRIC MOTORS AUTOMATICALLY, J. D. Leitch. *AIEE Transactions*, vol. 60, 1941, pp. 487–93.

5. FREQUENCY RELAYS AND MOTOR ACCELERATION FOR STEEL MILL CONTROL, D. C. Wright. *Proceedings*, Association of Iron and Steel Engineers, Pittsburgh, Pa., 1939, pp. 302–11.

A Series Method of Calculating Control-System Transient Response from the Frequency Response

DAVID V. STALLARD
ASSOCIATE MEMBER AIEE

THE methods of calculating the closed-loop frequency response (control ratio C/R) of a linear control system are well known and straightforward.[1] To calculate the transient response to a step reference input, which in some cases may be more important, is more difficult. Of the various methods[1–4] for calculating this transient response, the Laplace transformation is perhaps the most rigorous and also the most laborious, especially if the control ratio C/R has numerous poles. Partly because of this difficulty, the frequency response has predominated over the transient response in analysis and design. There is need for a simple, fast

method of calculating the step transient response, preferably from the frequency response.

This paper explains a simple method of calculating the step transient response from a convergent sine series formed directly from the control ratio. The technique is not wholly new, for in 1939 Bedford and Fredendall[5] published a sine series method for calculating the step transient response of video amplifiers. Using an independent but similar method, Tustin[6] published in 1952 a simpler sine series for calculating the transient response of a control system from its control ratio; but he did not show how to

make the series short or how to test it.

Tustin's method is the basis of this paper, which extends the method by showing a technique for forming a short but accurate sine series and for testing its convergence. The actual computation and plotting of the transient response is a simple routine which can be performed by an unskilled person.

Analysis of Transient Response

REPLACING STEP TRANSIENT RESPONSE BY AN EQUIVALENT PERIODIC RESPONSE

The fundamental principle of the sine series method of calculating the step

Fig. 1. A—Block diagram of a servo. B—Transient response to a step input

Fig. 2 (right). Equivalent square wave and periodic response
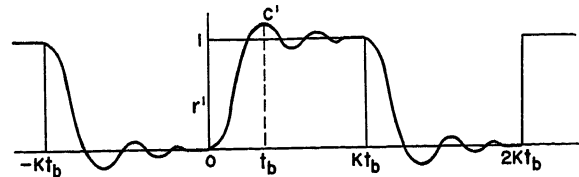
Fig. 3 (right). Periodic response to wave of doubled period

transient response of a control system is that the step input may be replaced by an equivalent square-wave input. To find an equivalent square wave, consider the block diagram of a typical control system and its response to a unit step input in Fig. 1. Because a slightly oscillatory step response is a typical and usually desirable response, it will be assumed here. The time for the first transient overshoot can be estimated as Floyd's[1] "build-up time t_b" (as discussed in Appendix I), which is given by

$$t_b \cong \frac{\pi}{\omega_0} \qquad (1)$$

where ω_0 is the cutoff frequency, at which C/R has a phase angle of -90 degrees. It turns out that in many control systems the build-up time to reach the peak overshoot is between $3/\omega_0$ and $4/\omega_0$ for moderate overshoots.

It is probable that in a well-designed system the step response will come to a vanishingly small velocity and error before time kt_b, where k is roughly 5. Hence, the input r and response c of the control system can be replaced without error over the time interval from 0 to kt_b by an equivalent square-wave reference input r' and a periodic response c' of Fig. 2. The period of this square-wave reference input r' is

$$T = 2kt_b \qquad (2)$$

and the fundamental frequency ω_1 is therefore

$$\omega_1 = 2\pi/T \qquad (3)$$

It is easily shown by elementary Fourier analysis[7] that the square-wave r' is the sum of the following infinite series

$$r' = \frac{1}{2} + \frac{2}{\pi} \sum_{n=1,3}^{\infty} \frac{\sin (n\omega_1 t)}{n} \qquad (4)$$

CALCULATING THE RESPONSE BY SUPERPOSITION

A series expression for the step response now follows readily, because the principle of superposition can be applied if the system is linear. The response to the square-wave input is the sum of the responses to the Fourier components taken separately. Thus, over the interval from 0 to kt_b, the transient response to a step input can be represented as a series of sine terms

$$c' = \frac{1}{2} + \frac{2}{\pi} \sum_{n=1,3}^{\infty} |W_n| \frac{\sin (n\omega_1 t + \underline{/W_n})}{n} \qquad (5)$$

where $|W_n|$ and $\underline{/W_n}$ denote the magnitude and phase angle respectively of the control ratio C/R at the frequency $n\omega_1$. This equation was published with different notation by Bedford and Fredendall[5] for calculating the transient response of video amplifiers to a step input of voltage; no explicit criterion for determining a suitable frequency ω_1 was given.

The infinite series of equation 5 for c'

will converge for two reasons: 1. the Fourier coefficients for r' in equation 4 vary inversely with frequency, the coefficient of the seventh harmonic term being only 0.091; and 2. above the cutoff frequency ω_0, the frequency response C/R diminishes rapidly with increasing frequency.

TUSTIN'S TIME SERIES

Tustin[6] has recently published two sine series for calculating control-system step response which are simpler and more convenient than the series expression of Bedford and Fredendall (equation 5). Although Tustin's derivation was independent of that of Bedford and Fredendall, the two derivations are similar, and it is interesting to show how Tustin's series can be developed from equation 5. The familiar trigonometric equation for the sine of the sum of two angles may be applied to equation 5 to give

$$c' = \frac{1}{2} + \frac{2}{\pi} \sum_{n=1,3}^{\infty} \times$$
$$\frac{|W_n| \cos \underline{/W_n} \sin (n\omega_1 t) + |W_n| \sin \underline{/W_n} \cos (n\omega_1 t)}{n} \qquad (6)$$

$$c' = \frac{1}{2} + \frac{2}{\pi} \sum_{n=1,3}^{\infty} \times$$
$$\frac{Re(W_n) \sin (n\omega_1 t) + Im(W_n) \cos (n\omega_1 t)}{n} \qquad (7)$$

where $Re(W_n)$ denotes the real part (in-phase component) of the control ratio C/R at the frequency $n\omega_1$, and $Im(W_n)$ denotes the imaginary part (quadrature component). Now suppose that the period $2kt_b$ of the square-wave input r' is doubled by doubling the value of k; then the response c' is zero from time $-kt_b/2$ to zero time, and is unity from $kt_b/2$ to kt_b; see Fig. 3. By this assumption, the response c' in equation 7 is zero if $-t$ is substituted for t

Table I. Re(Wₙ) for Instrument Servo

$n\omega_1$	Frequency, Radians per Second	$10 \log |W_n|$	Angle W_n, Degrees	$Re(W_n)$	$\dfrac{Re(W_n)}{n}$	$\dfrac{4}{\pi} \dfrac{Re(W_n)}{n}$
ω_1	10.9	0.33	-2.5	1.08	1.08	1.360
$3\omega_1$	32.7	0.88	-19.5	1.16	0.386	0.491
$5\omega_1$	54.5	1.0	-43	0.92	0.184	0.234
$7\omega_1$	76.2	0.82	-66	0.49	0.070	0.090
$9\omega_1$	98		-90	0	0	0
$11\omega_1$	120	-1.15	-110	-0.262	-0.024	-0.030
$13\omega_1$	142	-2.3	-121	-0.302	-0.023	-0.030
$15\omega_1$	163	-3.3	-131	-0.306	-0.020	-0.026
$17\omega_1$	185	-4.5	-139	-0.269	-0.016	-0.020

$$0 = \frac{1}{2} + \frac{2}{\pi} \sum_{n=1,3}^{\infty} \times$$

$$\frac{Re(W_n)[-\sin (n\omega_1 t)] + Im(W_n)\cos (n\omega_1 t)}{n}$$

$$(8)$$

Equations 7 and 8 may be added to give

$$c' = 1 + \frac{4}{\pi} \sum_{n=1,3}^{\infty} \frac{Im(W_n)\cos (n\omega_1 t)}{n} \qquad (9)$$

Alternatively, equation 8 may be subtracted from equation 7 to show that

$$c' = \frac{4}{\pi} \sum_{n=1,3}^{\infty} \frac{Re(W_n)\sin (n\omega_1 t)}{n} \qquad (10)$$

Tustin has derived equations 9 and 10 by a slightly different but equivalent treatment. Certainly equation 10 is simpler to calculate and plot versus time than the older series of Bedford and Fredendall (equation 5), and it may converge faster, even though the fundamental frequency ω_1 is necessarily halved.

However, Tustin did not offer an explicit criterion for choosing a fundamental frequency ω_1. If it is chosen too big, the series is inaccurate. If ω_1 is conservatively made quite small, then the series is long and the convenience of the method is lost.

CONVERGENCE

To extend the usefulness of Tustin's sine series for calculating step response, it is desirable to determine a good choice of ω_1 and to test the convergence of the resulting series. To choose ω_1, it may be assumed that the settling time of the transient response in Fig. 3 is less than 4.5 t_b. This assumption is discussed in Appendix B of reference 8, from which this paper is condensed. From Fig. 3 it is clear that the interval 4.5 t_b may be set equal to the quarter-period $kt_b/2$. Hence k may be equal to 9, and the period T is 18 t_b. The fundamental frequency is then

$$\omega_1 = \frac{2\pi}{T} = \frac{\pi}{9t_b} \cong \frac{\pi}{9\pi/\omega_0} = \frac{\omega_0}{9} \qquad (11)$$

using Floyd's approximate value π/ω_0 for t_b. The various harmonic frequencies are $\omega_0/3$, $5\omega_0/9$, $7\omega_0/9$, ω_0, $11\omega_0/9$, etc. The term in equation 10 at the ninth harmonic frequency ω_0 vanishes because at this frequency the phase angle of C/R is $-90°$ (by definition) and so the real part of C/R is zero. Terms at higher frequencies are quite small and the series converges rapidly.

As a preliminary check on the convergence and accuracy of the terms in the
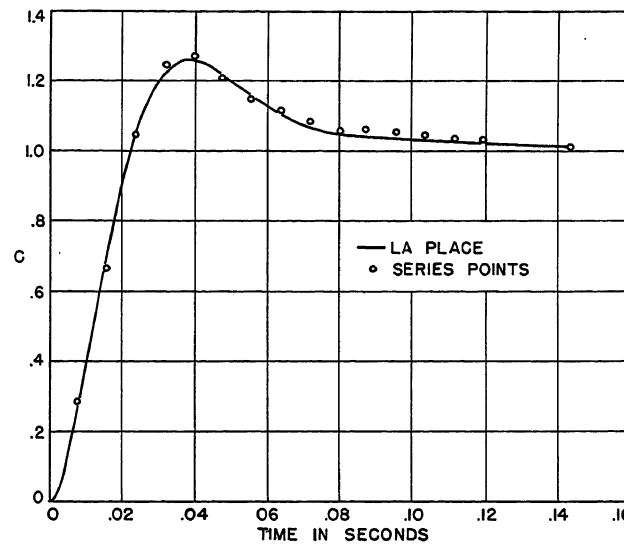
Fig. 4. Step response of instrument servo

series of equation 10, it can be evaluated at $kt_b/2$ by substituting 90 degrees for $\omega_1 t$. The response is

$$c' = \frac{4}{\pi}\left(Re(W_1) - \frac{Re(W_3)}{3} + \frac{Re(W_5)}{5} - \frac{Re(W_7)}{7}\cdots\right) \qquad (12)$$

and necessarily must be unity, if equation 10 is valid.

EXAMPLE: STEP RESPONSE OF AN INSTRUMENT SERVOMECHANISM

To illustrate the method, a simple example has been chosen. It is an actual instrument servomechanism[7] compensated by a viscous-coupled inertia damper, with a loop ratio which has been calculated as

$$\frac{B}{E} = G = \frac{1320(0.064s+1)}{s^2(0.00744s+1)} \qquad (13)$$

A plot of G on the log magnitude-angle diagram showed that the cutoff frequency ω_0 was 15.6 cycles per second (98.0 radians per second). Hence the build-up time was estimated as $\pi/98 = 0.033$ second, and the fundamental frequency was chosen as $98./9 = 10.9$ radians per second. From the plot of G on the log magnitude-angle diagram (not shown) the values of $Re(W_n)$ are calculated in Table I.

The convergence of the series was checked by substituting the coefficients of the right-hand column of Table I into equation 12. The sum was 1.019 instead of unity, but this discrepancy was probably within the range of computational errors.

The step transient response of the instrument servo was then calculated by equation 10 with the eight coefficients of Table I. The calculation was facilitated
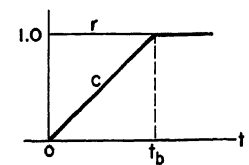

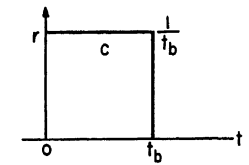**Fig. 5. Idealized step response**

Fig. 6. Unit impulse response

by a tabular form in which each column contained the terms for a given value of time; 5-degree intervals were chosen for $\omega_1 t$. The solution is plotted as the circled points in Fig. 4.

For purposes of checking and comparison, the step response was also rigorously calculated by the Laplace transform[2] method as

$$c = 1 + 0.305\epsilon^{-19.4t} + 1.588\epsilon^{-57.5t} \times \cos (76.4t - 214.8 \text{ degrees}) \qquad (14)$$

The Laplace solution is plotted as a smooth curve in Fig. 4. Only 4 of the 16 points calculated by the series solution differ by 3 per cent from the exact Laplace solution, the rest agreeing within 2 per cent. These discrepancies are due partly to computational errors. Also, the build-up time in Fig. 4 is about 0.037 second, instead of 0.033 second as estimated.

It is apparent that the solution calculated by the series is sufficiently accurate for engineering purposes. Furthermore, the coefficients of the series are quickly found from the plot of the loop ratio on the log magnitude-angle diagram, which is commonly used for design purposes anyway. After the coefficients of the series were checked by equation 12, it was possible to assign the computations to engineering assistants. The Laplace[2] solution is usually much more tedious to calculate than that in the example and generally cannot be assigned to engineering assistants.

TRANSIENT RESPONSE TO RAMP INPUTS

The transient response to a unit ramp input $(r=t)$ may be derived from equation 10 for the step response. It is easily shown that the response to a unit ramp input is the time integral of the response to a unit step input. Equation 10

may readily be integrated term by term. Computations may then proceed in much the same way as for the unit step input.

Conclusions

In the design of a linear control system, it is often desirable to calculate the step transient response, for this may be more important than the frequency response. By extending the series method of Tustin it is possible to calculate the step transient response almost directly from the control ratio C/R. The step transient response c is replaced by an equivalent periodic response c' to a square wave input r', whose period is more than four times the servo settling time. For example, the fundamental frequency ω_1 of the square wave is $\omega_0/9$, where ω_0 is the cutoff frequency, at which the control ratio C/R has a phase lag of 90 degrees. The response c' is given by

$$c' = \frac{4}{\pi} \sum_{n=1,3}^{\infty} \frac{Re(W_n)}{n} \sin(n\omega_1 t) \qquad (15)$$

where $Re(W_n)$ is the real part (in-phase component) of the control ratio C/R at the frequency $n\omega_1$, and is readily found from the log magnitude-angle diagram. Usually eight terms in the series will suffice to give adequate accuracy. The convergence and accuracy may be checked easily by letting $\omega_1 t$ equal 90 degrees and calculating the response; because of an original assumption in the derivation, the response should be unity. In the calculated example, the series solution agrees well with the Laplace transform solution and was much easier to calculate.

In the calculation of the step response, the series is quickly derived from the log magnitude-angle diagram. After the series has been checked for convergence and accuracy at $\omega_1 t = 90$ degrees, it can be assigned to assistants as a routine numerical calculation. Each calculation yields a point on the step response plot. There is no need for special curve fitting, or difficult solutions of poles, as in other meth-
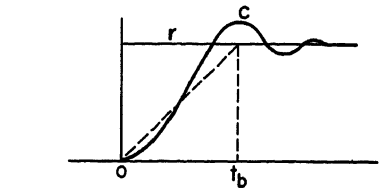
Fig. 7. Build-up time of an actual transient response

ods. The series method is direct, simple, and accurate. Although the discussion has been focused on feedback control systems, the series method of calculating step transient response can be applied to many other linear systems whose frequency response has been calculated.

Appendix I. Approximating the Build-up Time t_b

Floyd[1] has developed a transformation by which he "determines the system frequency response required to realize an idealized response to a unit step function, where the step is the time integral of the impulse." The idealized response is shown in Fig. 5. The corresponding response to a unit impulse is the time derivative of this idealized step response, and is shown in Fig. 6. For the sake of conciseness, this paper offers a mathematical treatment which is equivalent to Floyd's but is shorter. The control ratio C/R of the system can be found by dividing the Laplace transform of the unit impulse response by the transform of the unit impulse

$$\frac{C}{R} = \frac{L\text{ (response)}}{L\text{ (unit impulse)}} = \frac{\int_0^\infty r\epsilon^{-st}dt}{1} \qquad (16)$$

$$\frac{C}{R} = \int_0^{t_b} \frac{1}{t_b}\epsilon^{-st}dt = \frac{\epsilon^{-st_b}}{-st_b} + \frac{1}{st_b} \qquad (17)$$

To particularize the analysis for a sinusoidal frequency response, $j\omega$ may be substituted for s in equation 17

$$\frac{C}{R} = \frac{j}{\omega t_b}(\epsilon^{-j\omega t_b} - 1) \qquad (18)$$

Consider the case where C/R has a phase angle of -90 degrees. Evidently this can occur only at frequencies which make the exponential term equal to -1. The lowest frequency ω_0 at which this can happen is given by

$$-j\omega_0 t_b = -j\pi \qquad (19)$$

Hence the build-up time t_b is related uniquely to the cutoff frequency ω_0

$$t_b = \frac{\pi}{\omega_0} \qquad (20)$$

This is the same result which Floyd obtained with a different analysis beginning with an equation relating the real part of C/R to the impulse response.

Now, this concept of build-up time may be extended to the transient response of a physical system, as shown in Fig. 7. To a first approximation, the actual transient response can be replaced by an idealized response whose time t_b corresponds to the time for the first overshoot. Hence, the build-up time for the first peak overshoot of an actual transient can be estimated roughly as π/ω_0 from equation 12. On page 62 of reference 3, a graph of transient response for an elementary servomechanism shows that the build-up time is about $3.3/\omega_0$ if the peak overshoot is 25 per cent. Hence it is intuitively concluded that the build-up time can be approximated as π/ω_0 or $3.3/\omega_0$ for reasonable overshoots.

References

1. PRINCIPLES OF SERVOMECHANISMS (book), G. S. Brown, D. P. Campbell. John Wiley and Sons, Inc., New York, N. Y., 1948. (Particular reference is made to chapter 11 which is based on a thesis by G. F. Floyd.)

2. TRANSIENTS IN LINEAR SYSTEMS (book), M. F. Gardner, J. L. Barnes. *Ibid.*, 1942.

3. SERVOMECHANISMS AND REGULATING SYSTEM DESIGN (book), H. Chestnut, R. W. Mayer. *Ibid.*, 1951.

4. THEORY OF SERVOMECHANISMS (book), H. M. James, N. B. Nichols, R. S. Phillips. McGraw-Hill Book Company, New York, N. Y., 1947.

5. TRANSIENT RESPONSE OF MULTISTAGE VIDEO-FREQUENCY AMPLIFIERS, A. V. Bedford, G. L. Fredendall. *Proceedings,* Institute of Radio Engineers, New York, N. Y., vol. 27, 1939, p. 277.

6. D.C. MACHINES FOR CONTROL SYSTEMS (book), A. Tustin. MacMillan Company, New York. N. Y., 1952.

7. A DAMPER STABILIZED SERVO DATA REPEATER, J. E. Ward. *Proceedings of the 1954 National Telemetering and Remote Control Conference,* Institute of Radio Engineers, New York, N. Y., Jan. 1955.

8. A SERIES METHOD OF CALCULATING SERVOMECHANISM TRANSIENT RESPONSE FROM THE FREQUENCY RESPONSE, D. V. Stallard. *Report No. 7138-R-2,* Massachusetts Institute of Technology Servomechanisms Laboratory, Cambridge, Mass., Nov. 16, 1953.

9. METHODS OF ADVANCED CALCULUS (book), P. Franklin. McGraw-Hill Book Company, New York, N. Y., 1944.

Analysis of Errors in Sampled-Data Feedback Systems

JACK SKLANSKY
NONMEMBER AIEE

J. R. RAGAZZINI
MEMBER AIEE

Synopsis: The system error in sampled-data feedback systems resulting from the application of a test function is an important design parameter. The "system error" is defined as the difference between the actual output of the system and the desired output. There are two components which contribute to this error: one, called "organic error," is brought about by the system energy storages, and the other, called "ripple," is brought about by the sampling process. Formulation of the system error is obtained by the application of the Laplace and z-transformations. Both the system error time function in intersampling times and the rms system error are formulated in terms of the system parameters. Illustrative examples demonstrate the theory.

A N IMPORTANT class of feedback systems is characterized by the fact that the data appear in sampled form in one or more places in the system. A common prototype of a sampled-data system is shown schematically in Fig. 1, where the error signal appears in sampled form. The sampler is shown as a synchronous mechanical switch whose period between samples is T seconds. The hold circuit following the switch is used to convert the pulse sequence into a continuous function more closely resembling the original error function. This paper will consider hold circuits which are implemented either with linear networks or with the standard clamp or boxcar circuit.

As in the case of continuous systems, the sampled-data system will experience a system error which is the result of the imperfect response of the system to an applied input. One component, which will be referred to as the organic error, is introduced by the lags or leads of the continuous part of the system. In addition, there is another component of system error, usually called ripple, which is introduced by the sampler. In the steady state, this ripple contains only those frequency components which are at sampling frequency or its harmonics. Since the sampling frequencies are generally high with respect to the passband of the system, ripple effects originating in the sampler are generally suppressed by the forward transmission function (including the hold) but, except for certain cases, never perfectly.

Despite the normally small amount of ripple in the output of a sampled-data system, it does represent a degradation in performance which may be serious in certain applications. For instance, in precise control systems where the output must follow an input very closely, ripple effects may cause the performance of the system to fall outside acceptable tolerances. In systems where the controlled variable is at a high power level, ripple causes an unnecessary loss of power in the actuator. In systems employing gearing, ripple will cause a hammering of the gears, which will increase wear and backlash. For systems which include a human operator, ripple will cause unnecessary control movements on his part and thereby induce fatigue. For process control, ripple causes a variation in the quality of the product. It is evident, therefore, that both organic and ripple components of the system error require investigation. Earlier work has been done on the subject,[1-4] but this paper presents novel approaches to the problem which lead to results useful to the designer.

The problem of obtaining mathematical descriptions for ripple is not a simple one. An expedient formulation will be developed through the use of z-transform[1] and ordinary Laplace transform techniques. (The z-transformation is the application of the Laplace transform to pulsed data. Because this obtains functions that are rational in ϵ^{sT} this variable is replaced by z and the resulting functions are called z-transforms.) During transient disturbances, the ripple component of system error is generally small and often meaningless. In the steady state, on the other hand, the ripple may be the only component of the system error.

To establish the significance of system error, reference is made to Fig. 2. Plotted here is a time function $c_d(t)$ which represents the desired response resulting from the application of an input and the time function $c(t)$ representing the actual response. The difference between these two functions represents the system error $\epsilon(t)$, part of which is organic and the remainder is ripple. Stated mathematically

$$\epsilon(t) \triangleq c_d(t) - c(t) \tag{1}$$

To obtain the system error, it is necessary to place some specification on the desired response. For instance, in a unity feedback system the desired response is generally taken equal to the input. More difficult to ascertain in the case of sampled-data systems is the actual response function $c(t)$. The material which follows is largely concerned with the determination of $c(t)$ for the sampled-data feedback control system.

General Formulations

As mentioned previously, the combination of z-transform and ordinary Laplace transform techniques can be employed to obtain the time function for the output. For the system of Fig. 1, it has been shown[1] that the Laplace transform of the output time function $c(t)$ is given by

$$C(s) = \frac{G(s)}{1 + GH^*(z)R^*(z)} \tag{2}$$

where

$z \triangleq$ auxiliary variable ϵ^{sT}
$GH^*(z) \triangleq$ pulsed loop transfer function
$R^*(z) \triangleq z$-transform of the input function

The central problem encountered is that of readily obtaining the inverse of this transform. While this is not a simple task, procedures and methods given later will result in a workable solution.

As a basis for derivation, the system shown in Fig. 3(A) will be considered. The hold shown here is a simple clamp
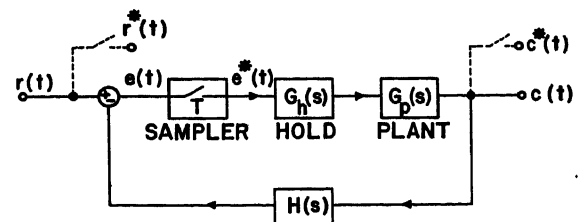
Fig. 1. Typical error-sampled feedback system. T is the sampling period

circuit (or zero-order hold). It is emphasized that the analysis applies equally well to holds in the form of low-pass networks. A little reflection will show that the system block diagram can be rearranged as shown in Fig. 3(B). Here the integrating action of the hold is absorbed in the plant and the operation of subtracting the present pulse from the one immediately preceding is distinguished in a separate block.

For physical systems, the transfer function $G_p(s)$ has more poles than zeros and, furthermore, the poles will generally be simple. Also, the transfer functions of most practical control systems contain a pole at the origin of the complex plane, thus giving $G_p(s)/s$ a double pole at the origin. As a consequence of these properties, the partial fraction expansion of $G_p(s)/s$ will in general appear as follows

$$\frac{G_p(s)}{s} = \frac{A_1}{s^2} + \frac{A_2}{s} + \frac{A_3}{s+a_3} + \frac{A_4}{a+a_3} +$$

$$\ldots = \sum_m G_m(s) \quad (3)$$

Schematically, the block diagram of the system can be redrawn in the form shown in Fig. 4 where it is seen that the forward transmission can be resolved into a number of simple parallel paths corresponding to each term of equation 3. Applying the z-transform technique to obtain the z-transform of the pulsed output of each parallel channel, it is possible to relate the output of the mth channel, $c_m{}^*(t)$, to the input to that channel, $e_1{}^*(t)$, by the relation

$$C_m{}^*(z) = E_1{}^*(z) G_m{}^*(z)$$

$$= \frac{(1-z^{-1})R^*(z)}{1+GH^*(z)} G_m{}^*(z) \quad (4)$$

where $G_m{}^*(z)$ is the pulsed transfer function of the mth path. For instance, for the first parallel path shown in Fig. 4 the function $G_1{}^*(z)$ used in equation 4 is the z-transform corresponding to a transfer function A_1/s^2.

By inverting the z-transforms† for each path, it is a relatively simple matter to

---

†The general inversion equation is $c(nT) = 1/(2\pi j) \int_\Gamma z^{n-1} C^*(z) dz$, where Γ is a circle centered about the origin of the z-plane and enclosing all the finite poles of $C^*(z)$. See Stone,[5] p. 217, and Smith, Lawden, and Bailey,[6] p. 380.
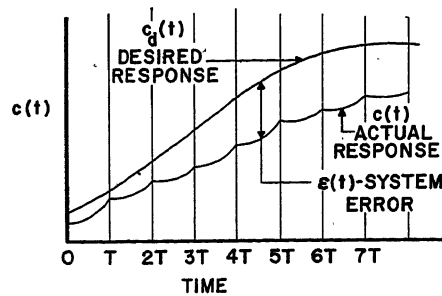
Fig. 2. Curve showing significance of system error

obtain the magnitude of the pulses constituting the output pulse train for that particular channel. The pulse at the nth sampling instant, $c_m(nT)$, determines the continuous time function for the output of the mth channel in the interval which follows. Thus for a channel such as the third or fourth in Fig. 3, whose transfer function is $A_m/(s+a_m)$, this time function is

$$c_m(t) = c_m(nT) e^{-a_m(t-nT)} \quad (5)$$

The expression is simple so long as the initial value of the transient, $c_m(nT)$, is available. This has been readily obtained by the use of the z-transform in equation 4.

The only component of the forward transfer function for a particular channel which may give difficulty is the one having a double pole at the origin. For instance, the first channel in Fig. 3 has a transfer function A_1/s^2. The time function between the nth and $(n+1)st$ instants, $c_1(t)$, is of the form

$$c_1(t) = a + b(t-nT) \quad (6)$$

The constants a and b may be evaluated by noting that a, the value at the beginning of the interval, is $c_1(nT)$, while the slope b is the product of the error pulse $e_1(t)$ and the constant A_1. The error pulse is readily obtainable from the inversion of $E_1{}^*(z)$.

It should be observed that the technique outlined here is applicable equally well in cases where a clamp circuit is not included. Should a low-pass filter replace this element, the only change in the technique outlined in the foregoing is that the term $(1-z^{-1})$ is omitted from equation 4 and that the factor $1/s$ is

omitted from the transfer function $G_p(s)/s$.

The complete continuous output function $c(t)$ for any interval between nT and $(n+1)T$ is the sum of the output time functions for each parallel channel. Thus

$$c(t) = c_1(t) + c_2(t) + \ldots + c_m(t) \quad (7)$$

where $nT \leqslant t < (n+1)T$.

The total system error $\epsilon(t)$ is obtained by substitution of $c(t)$ in equation 1. In the steady state the system error is generally readily separable into organic and ripple components. This is not always the case during transient or for more complex input functions. In any case, it is not of paramount importance to separate the system error into components since its total magnitude determines the performance of the feedback system under study. What is important are the contributions to the system error by the system parameters. This information will be of value to the designer in meeting error specifications and will be available in the expressions for system error obtained by the foregoing procedure.

ILLUSTRATIVE EXAMPLE

To illustrate the procedure which has been outlined, its application to the simple feedback system shown in Fig. 5 is described. In this system a lag network, rather than a clamp circuit, serves as the hold. The actuator or plant is taken as a perfect integrator. The feedback transfer function is unity and the system is error-sampled. The problem is to find the output time function for the interval of time between the nth and $(n+1)st$ sampling instant. The forward transfer function $G(s)$ is

$$G(s) = \frac{1}{s(s+1)} \quad (8)$$

This can be expanded into partial fractions

$$G(s) = G_1(s) + G_2(s) = \frac{1}{s} - \frac{1}{s+1} \quad (9)$$

Using the available tables of z-transforms[1] the z-transforms of the forward transfer function and its components are obtained
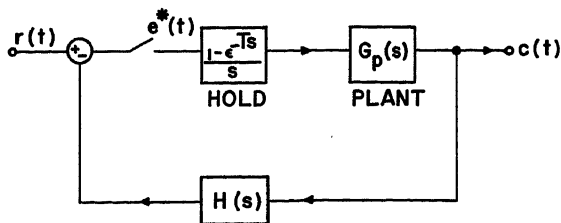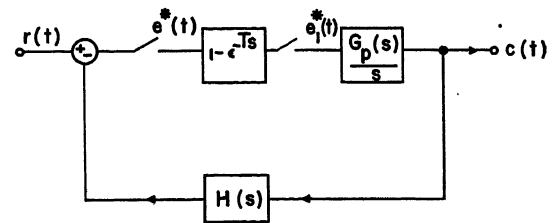
$$G_1{}^*(z) = \frac{1}{1-z^{-1}} \quad (10)$$

Fig. 3. System used for derivation of system error

A (left)—Actual system
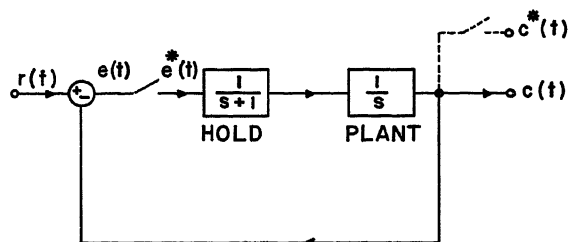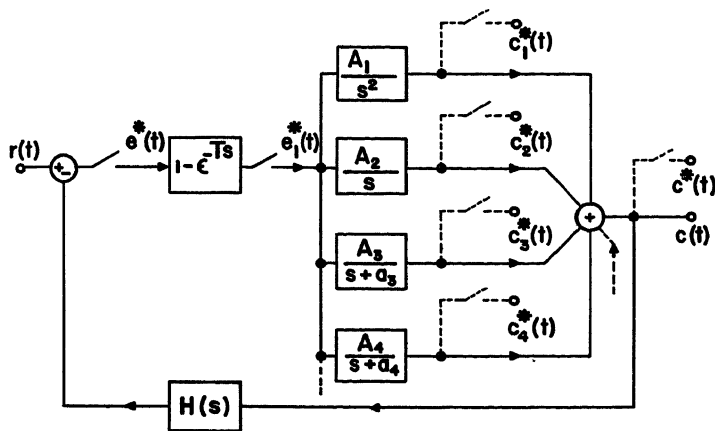B (right)—Rearranged system

Fig. 4 (left). Equivalent system for determination of system error

Fig. 5 (above). Feedback system used in illustrative examples

$$G_2^*(z) = -\frac{1}{1-\epsilon^{-T}z^{-1}} \tag{11}$$

$$G^*(z) = \frac{(1-\epsilon^{-T})z^{-1}}{(1-z^{-1})(1-\epsilon^{-T}z^{-1})} \tag{12}$$

The components in each channel are obtained by the general equation 4 except that the hold transfer function is omitted. In addition, it is assumed that the test input $r(t)$ is a unit step function, making $R^*(z)$ equal to $1/(1-z^{-1})$. Thus, substituting in equation 4 there results

$$C_1^*(z) = \frac{1}{1-z^{-1}} + \frac{\epsilon^{-T}z^{-1}}{1-2\epsilon^{-T}z^{-1}+\epsilon^{-T}z^{-2}} \tag{13}$$

$$C_2^*(z) = -\frac{1}{1-2\epsilon^{-T}z^{-1}+\epsilon^{-T}z^{-2}} \tag{14}$$

If, for purposes of illustration, it is assumed that T is chosen equal to $\ln 2$, so that $\epsilon^{-T}=1/2$, then inversion of equations 13 and 14 will yield

$$c_1(nT) = 1+2^{-n/2}\sin\frac{n\pi}{4} \tag{15}$$

and

$$c_2(nT) = -2^{-n/2}\left(\cos\frac{n\pi}{4}+\sin\frac{n\pi}{4}\right) \tag{16}$$

The time function in the interval from nT to $(n+1)T$ is obtained by the use of equation 5

$$c(t) = c_1(nT)+c_2(nT)\epsilon^{-(t-nT)} \tag{17}$$

which, upon substitution of the results of equations 15 and 16, becomes

$$c(t) = 1-2^{-n/2}\cos\frac{n\pi}{4}\epsilon^{-(t-nT)} + 2^{-n/2}\sin\frac{n\pi}{4}(1-\epsilon^{-(t-nT)}) \tag{18}$$

To find the system error, the desired output $c_d(t)$ must be specified. Since the feedback is unity, the desired output is taken equal to the input $r(t)$. Hence

$$\epsilon(t) = r(t)-c(t) \tag{19}$$

Since $r(t)$ is unity, the system error in the interval $nT \le t < (n+1)T$ becomes

$$\epsilon(t) = 2^{-n/2}\left[\cos\frac{n\pi}{4}\epsilon^{-(t-nT)} - \sin\frac{n\pi}{4}(1-\epsilon^{-(t-nT)})\right] \tag{20}$$

A measure of performance throughout all sampling intervals is obtained in this manner.

Rms System Error

In the previous section the output function $c(t)$ was obtained and the difference between this function and a desired output function was called the system error. This error can be evaluated both during transients and in the steady state. The results are particularly valuable when it is desired to ascertain whether the error exceeds some maximum allowable tolerance. Often, however, an average performance of the system is significant and a designation of rms error is desirable. As in all averaging procedures, the function which is being averaged should have constant or asymptotic properties over all time if the result is to be significant. For this reason, in obtaining an expression for the rms error, only steady-state operation is considered.

A general result can be obtained if the input test function $r(t)$ is taken as an exponential or sum of exponentials of the form

$$r(t) = \epsilon^{\lambda t} \tag{21}$$

where λ is complex for the general case. (In particular, when $\lambda=j\omega$, the equation to be derived for the per-cent mean square ripple applies to a sinusoidal input of frequency ω as well as to $\epsilon^{j\omega t}$, and this property can in turn be generalized to periodic and random inputs. This is demonstrated in Appendix II.) For use in equation 2, which gives the Laplace transform of the output, the transform of the sampled function $r^*(t)$ must be used. It has been shown[1,2] that for the input just given

$$R^*(z) = \frac{1}{T}\sum_{n=-\infty}^{\infty} R(s+jn\omega_0)$$
$$= \frac{1}{T}\sum_{n=-\infty}^{\infty}\frac{1}{s-(\lambda+jn\omega_0)} \tag{22}$$

where

$$R(s)\underset{=}{\Delta}\mathcal{L}\epsilon^{\lambda t}=\frac{1}{s-\lambda}, \quad \omega_0\underset{=}{\Delta}\frac{2\pi}{T}$$

Equation 22 is substituted in equation 2. The steady-state output $c_{ss}(t)$ is obtained by evaluating the residues of the function $C(s)$ only at the poles of $R^*(z)$. Thus, the expression for the steady-state output time function becomes

$$c_{ss}(t) = \frac{1}{T}\sum_{n=-\infty}^{\infty}\frac{G(\lambda+jn\omega_0)\epsilon^{(\lambda+jn\omega_0)t}}{1+GH^*(\epsilon^{(\lambda+jn\omega_0)T})}$$
$$= \frac{\left(\sum_{n=-\infty}^{\infty}G(\lambda+jn\omega_0)\epsilon^{jn\omega_0 t}\right)\epsilon^{\lambda t}}{T[1+GH^*(\epsilon^{\lambda t})]}, \tag{23}$$

since $GH^*(z)$ is periodic in $j\omega$.

It is of interest to note that a sampled-data system is linear but time-variant. L. A. Zadeh has shown[7] that such a system can be characterized by a system function containing t as a parameter; for the sampled-data system considered here, the system function is

$$K(s;t)\underset{=}{\Delta}\left.\frac{c_{ss}(t)}{\epsilon^{\lambda t}}\right|_{\lambda=s}=\frac{\sum_{n=-\infty}^{\infty}G(s+jn\omega_0)\epsilon^{jn\omega_0 t}}{T[1+GH^*(\epsilon^{sT})]}$$

The desired function $c_d(t)$ is taken to be the signal component of the steady-state output, i.e., the component of complex frequency, λ, obtained from equation 23 by setting $n=0$. Thus the desired function is

$$c_d(t) = \frac{G(\lambda)\epsilon^{\lambda t}}{T[1+GH^*(\epsilon^{\lambda T})]} \tag{24}$$

and the system error is the difference between the complete steady-state response $c_{ss}(t)$ and the desired response $c_d(t)$. This can be considered to be the steady-state

noise, often called ripple, due to the operation of the sampler.

If the mean-square system error is to be obtained, the following expression must be evaluated

$$\overline{|\epsilon(t)|^2} = \overline{|c_{ss}(t) - c_d(t)|^2}, \tag{25}$$

where the bar signifies a time average. In order that the mean be finite and non-zero, it is necessary that the real part of λ be zero, in which case only sinusoidal or d-c inputs have significance. For this condition, it is also noted that the desired output function $c_d(t)$ and the side-band components of $c_{ss}(t)$ are orthogonal, so that the mean of the cross product of these terms is zero. Thus the mean square of the system error becomes more simply

$$\overline{|\epsilon(t)|^2} = \overline{|c_{ss}(t)|^2} - \overline{|c_d(t)|^2} \tag{26}$$

The problem resolves itself to that of obtaining the mean-square value of expressions such as 23. To facilitate this, a general property of z-transforms will be utilized. This property is stated by

$$\mathcal{Z}[G(s)G(-s)]_{z=1} = \frac{1}{T} \sum_{n=-\infty}^{\infty} |G(jn\omega_0)|^2 \tag{27}$$

where $\mathcal{Z}[G(s)G(-s)]$ is the z-transform of the impulsive response of the system whose transfer function is $G(s)G(-s)$. The condition that z is equal to unity is equivalent to stating that s is equal to zero. The validity of equation 27 is seen by noting from references 1 and 2 that

$$\mathcal{Z}[G(s)G(-s)]$$
$$= \frac{1}{T} \sum_{n=-\infty}^{\infty} G(s+jn\omega_0)G(-s-jn\omega_0) \tag{28}$$

and setting s equal to zero. This property is extended slightly by introducing the complex frequency λ

$$\mathcal{Z}[G(s+\lambda)G(-s+\bar{\lambda})]_{z=1}$$
$$= \frac{1}{T} \sum_{n=-\infty}^{\infty} |G(\lambda+jn\omega_0)|^2 \tag{29}$$

where $\bar{\lambda}$ is the conjugate of λ. This property is applied to equation 23, where it is recognized that to obtain the mean square requires the summation of terms similar to those in equation 29. To facilitate notation, a function $P(s,\lambda)$ is introduced which is defined as

$$P(s,\lambda) \triangleq G(s+\lambda)G(-s+\bar{\lambda}) \tag{30}$$

Also

$$P^*(z,\lambda) \triangleq \mathcal{Z}[P(s,\lambda)] \tag{31}$$

The mean-square value of the steady-

state output is the sum of the squares of its frequency components. Hence with the application of equation 29 to 23 the mean-square value of the steady-state output is obtained as

$$\overline{|c_{ss}(t)|^2} = \frac{\sum_{n=-\infty}^{\infty} |G(\lambda+jn\omega_0)|^2}{T^2|1+GH^*(\epsilon^{\lambda T})|^2}$$
$$= \frac{P^*(z,\lambda)|_{z=1}}{T|1+GH^*(\epsilon^{\lambda T})|^2} \tag{32}$$

and the mean-square value of the desired function is

$$\overline{|c_d(t)|^2} = \frac{|G(\lambda)|^2}{T^2|1+GH^*(\epsilon^{\lambda T})|^2} \tag{33}$$

(A more detailed derivation of the second member of equation 32 is given in Appendix I.) Thus the mean-square value of the system error is

$$\overline{|\epsilon(t)|^2} \triangleq \overline{\epsilon^2(\lambda)} = \frac{TP^*(z,\lambda)|_{z=1} - |G(\lambda)|^2}{T^2|1+GH^*(\epsilon^{\lambda T})|^2} \tag{34}$$

where the mean square is seen to be a function of λ and is so written. This is the working equation for system error for error-sampled feedback systems.

If it is desired to take into consideration input functions in which λ has a positive real part, a modification of the mean-square system error can be introduced. By taking a weighted mean in which the weighting factor is $\exp\{-2[Re(\lambda)]t\}$, the squared functions being averaged are made to conform with the condition that the effective real part of λ in the exponentials in equations 23 and 24 be zero. Physically the use of this weighting factor is seen to be justifiable when it is considered that the error of the system when subjected to a rising exponential input would diverge and not possess a finite mean. Yet this specification of error has value because it is possible that a system be subjected to a disturbance which can be so characterized for a finite time, and the minimization of a steady-state weighted error would then have physical significance. The expression for error for this case is also given by equation 34.

The system error can also be expressed as a percentage of the desired output. If this is taken as the ratio of the mean-square system error, as given in equation 34, to the mean-square desired output, as given in equation 33, a much simplified result is obtained

$$\overline{\epsilon_p^2(\lambda)} \triangleq \frac{\overline{\epsilon^2(\lambda)}}{\overline{|c_d(t)|^2}} = \frac{TP^*(z,\lambda)|_{z=1}}{|G(\lambda)|^2} - 1 \tag{35}$$

Note that $[\overline{\epsilon_p^2(\lambda)}]^{1/2}$ may be called "ripple factor" in analogy with the terminology for power rectifiers. As might be expected from physical considerations, the

per-cent system error depends only on the forward transfer function $G(s)$. It is independent of both the relative level of forward gain and the feedback transfer function $H(s)$. This is a significant property which is of value in the design of system in which system error is a stringent requirement.

A more convenient expression for $\overline{\epsilon_p^2(\lambda)}$ can be given when λ is purely imaginary. Note that

$$P^*(z,j\omega)|_{z=1}$$
$$= \frac{1}{T} \sum_{n=-\infty}^{\infty} [G(s+j\omega+jn\omega_0)G(-s-$$
$$j\omega-jn\omega_0)]|_{s=0}$$
$$= \frac{1}{T} \sum_{n=-\infty}^{\infty} [G(j\omega+jn\omega_0)G(-j\omega-jn\omega_0)]$$
$$= \mathcal{Z}[G(s)G(-s)]_{z=\epsilon^{j\omega T}} \tag{36}$$

Thus, if one defines

$$Q(s) \triangleq G(s)G(-s) \tag{37}$$

then

$$P^*(z,j\omega)|_{z=1} = Q^*(z)|_{z=\epsilon^{j\omega t}} = Q^*(\epsilon^{j\omega t}) \tag{38}$$

Hence, when $\lambda = j\omega$, equation 35 becomes

$$\overline{\epsilon_p^2(j\omega)} = \frac{TQ^*(\epsilon^{j\omega T})}{|G(j\omega)|^2} - 1 \tag{39}$$

where $Q^*(\epsilon^{j\omega T})$ is the z-transform of $G(s)G(-s)$, with z replaced by $\epsilon^{j\omega T}$.

The per-cent mean-square ripple specification is a good figure of demerit to be minimized in a system design.

Expression 34 has the further advantage that statistical minimization procedures can be applied for inputs which are not expressed by a single exponential but by a spectral density. If the input signal spectral density is taken as $S_{rr}(\omega)$, then the mean-square value of the system error becomes

$$\overline{\epsilon^2(t)} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \overline{\epsilon^2(j\omega)} S_{rr}(\omega) d\omega \tag{40}$$

This follows from the fact that $\overline{\epsilon^2(\lambda)}$ becomes, roughly speaking, the square of the absolute value of the Fourier transform of the system error when λ is replaced by $j\omega$. (A more rigorous derivation of equation 40 is given in Appendix II.) The mean square of the desired output is similarly obtained

$$\overline{c_d^2(t)} = \int_{-\infty}^{\infty} |K_d(j\omega)|^2 S_{rr}(\omega) d\omega \tag{41}$$

where

$$K_d(j\omega) \triangleq \frac{G(j\omega)}{T[1+GH^*(\epsilon^{j\omega T})]}$$

A figure of demerit which can be used in

minimization procedures is the ratio of the mean-square system error to the mean-square desired output

$$F \triangleq \frac{\overline{\epsilon^2}}{\overline{c_d{}^2}} \qquad (42)$$

Often the integrals given in equations 40 and 41 will need to be evaluated only over a frequency band up to the sampling frequency or one or two multiples. This follows from the fact that the forward transmission rarely has bandwidth sufficient to pass a signal whose frequency is comparable to that of the sampler.

Generally, it will be difficult to evaluate integrals of the type given in equations 40 and 41 analytically, and numerical methods will have to be used. One such method consists in approximating the spectral density of the input signal by means of a series of equivalent sinusoidal terms, obtaining the mean-square error resulting from these term by term, and adding the result. Nevertheless, techniques for evaluating the integrals should be found so that analytical procedures which will statistically minimize the mean-square system error may be devised.

ILLUSTRATIVE EXAMPLE

To illustrate the mean-square method of specifying the system error, the system shown in Fig. 5 will be used. It is assumed that a sinusoidal input is applied and that the system is in a steady state. In this case the complex frequency λ will be taken as $j\omega$. The first step will be to obtain $P(s, j\omega)$, as defined in equation 30. Thus, for the $G(s)$ of Fig. 5

$$P(s, j\omega) = \frac{1}{(s+j\omega)^2(s+j\omega+1)(s+j\omega-1)} \qquad (43)$$

which upon expansion into partial fractions becomes

$$P(s, j\omega) = -\frac{1}{(s+j\omega)^2} + \frac{1}{2}\left[\frac{1}{s+j\omega-1} - \frac{1}{s+j\omega+1}\right] \qquad (44)$$

Using the tables of z-transforms in reference 1, and noting that $\mathcal{Z}[-1/(s+a)^2]$ is simply $[\partial/(\partial a)]\mathcal{Z}[1/(s+a)]$, the pulsed transform of $P(s, j\omega)$ becomes

$$P^*(z, j\omega) = \frac{Tz^{-1}\epsilon^{-j\omega T}}{(1-\epsilon^{-j\omega T}z^{-1})^2} + \frac{z^{-1}\epsilon^{-j\omega T}\sinh T}{[1-\epsilon^{-(j\omega+1)T}z^{-1}][1-\epsilon^{-(j\omega-1)T}z^{-1}]} \qquad (45)$$

(It will be noted that, since the input is sinusoidal, the use of equation 36 in obtaining 45 would have been somewhat more convenient, but for illustrative purposes the more general technique has been used.)

Setting z equal to unity

$$P^*(1, j\omega) = \frac{T}{4\sin^2\left(\frac{\omega T}{2}\right)} - \frac{\sinh T}{2(\cosh T - \cos \omega T)} \qquad (46)$$

Applying equation 35 to 46, with λ replaced by $j\omega$, the per-cent system error is given by

$$\overline{\epsilon_p{}^2}(j\omega) = \frac{TP^*(1, j\omega)}{|G(j\omega)|^2} - 1 \qquad (47)$$

which, upon substituting the results from equation 46, yields

$$\overline{\epsilon_p{}^2}(j\omega) = T\omega^2(\omega^2+1)\left[\frac{T}{4\sin^2\left(\frac{\omega T}{2}\right)} - \frac{\sinh T}{2(\cosh T - \cos \omega T)}\right] - 1 \qquad (48)$$

Letting T be equal to ln 2, as before, it is possible to calculate the system per-cent mean-square error as a function of input frequency. This requires the use of a desk calculator in view of the fact that relatively small differences of large numbers are involved. The per-cent rms system error is shown in Fig. 6. It is seen that the per-cent rms system error increases as the signal frequency increases. As a matter of fact, the rms ripple introduced by the sampler is, for this example, almost equal to the rms component of the output at an input frequency of half the sampling frequency. It is seen that to keep the ripple down below 10 per cent for a system of the type being considered, it is necessary to adjust the sampling frequency to be about 5 times the maximum input frequency.

By use of a better hold, or "desampling filter" as it is sometimes called, the per-cent ripple can be reduced. The techniques which are used to obtain the result with other forms of holds are the same as those outlined in the foregoing, with the
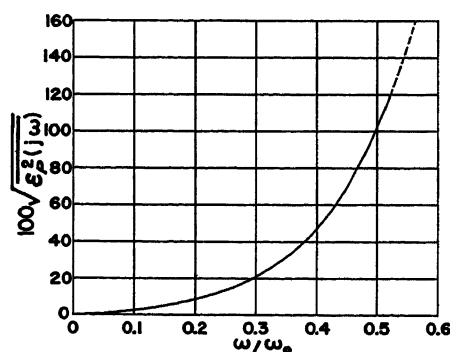
Fig. 6. Per-cent rms system error as a function of input frequency; $\omega_0 \triangleq 2\pi \times$ sampling frequency

possible exception that for some types of holds, such as a clamp or zero-order hold, rational functions of ϵ^{sT} are encountered in the forward transfer function.

Conclusions

The problem of obtaining the output function of a sampled-data feedback system between sampling instants can be readily solved by a combination of techniques involving continuous and pulsed Laplace transforms, the latter being known as z-transforms. The system error of a sampled-data system can be evaluated throughout any chosen interval, so that adherence to specifications in this regard can be readily ascertained at critical intervals. The effect of system parameters can be observed in view of the relatively straightforward procedures which are used. These procedures are valuable for the analysis of the transient as well as the steady-state response of the system.

Some classes of signal are not sudden disturbances, but are, rather, continuous time functions which can be described either in terms of a Fourier series or in terms of a spectral density. In this case the mean square of the steady-state system error is more meaningful. Here mean-square system error evaluations are made and the effect of system parameters observed. Unfortunately, in those cases where the signal is expressed in terms of a spectral density, analytical techniques for the evaluation of the mean-square system error are not available and this evaluation must be approximated by graphical or similar means. The techniques for evaluation of system error are developed and the performance of the system during intersample intervals can be specified. These techniques are valuable additions to the method employing the z-transform, which gives the output only at sampling instants.

Appendix I. Expression for the Mean-Square of the Response to $\epsilon^{\lambda t}$

Consider the expression for $c_{ss}(t)$ as given by equation 23. Note that

$$c_{ss}(t) = K(\lambda; t)\epsilon^{\lambda t} \qquad (49)$$

where $K(\lambda; t)$, the system function with s replaced by λ, is given by

$$K(\lambda; t) = \frac{\sum_{n=-\infty}^{\infty} [G(\lambda+jn\omega_0)\epsilon^{jn\omega_0 t}]}{T[1+GH^*(\epsilon^{\lambda t})]} \qquad (50)$$

Note that when $\lambda = j\omega$

$$\overline{|c_{ss}(t)|^2} = \overline{|K(j\omega;t)|^2} \quad (51)$$

Equation 50 can be extended to the case where λ has a positive real part by weighting the averaging of $|c_{ss}(t)|^2$ by the factor $\exp\{-2[Re(\lambda)]t\}$. One then obtains

$$\overline{|c_{ss}(t)\epsilon^{-\{Re(\lambda)\}t}|^2} = \overline{|K(\lambda;t)|^2}$$

If so desired, this may be written

$$\overline{|c_{ss}(t)|^2} = \overline{|K(\lambda;t)|^2} \quad (51A)$$

where it is understood that the left member is a weighted mean whenever $Re(\lambda)>0$. Now, squaring the absolute value of equation 50 gives

$$|K(\lambda;t)|^2 = \frac{1}{T^2|1+GH^*(\epsilon^{\lambda T})|^2} \times$$
$$\left\{ \sum_{n=-\infty}^{\infty} |G(\lambda+jn\omega_0)|^2 + \right.$$
$$\sum_{\substack{n=-\infty \\ n\neq m}}^{\infty} \sum_{m=-\infty}^{\infty} [G(\lambda+jn\omega_0) \times$$
$$\left. G(\bar\lambda-jm\omega_0)\epsilon^{j(n-m)\omega_0 t}] \right\} \quad (52)$$

where $\bar\lambda \triangleq$ conjugate of λ.

The mean of the second summation is zero. Hence, averaging equation 52 over t, one obtains

$$\overline{|K(\lambda;t)|^2} = \frac{\sum_{n=-\infty}^{\infty} |G(\lambda+jn\omega_0)|^2}{T^2|1+GH^*(\epsilon^{\lambda T})|^2} \quad (53)$$

Combining this with equation 51 (A), it is seen that when the input is $\epsilon^{\lambda t}$, $Re(\lambda)\geqslant 0$, then

$$\overline{|c_{ss}(t)|^2} = \frac{\sum_{n=-\infty}^{\infty} |G(\lambda+jn\omega_0)|^2}{T^2|1+GH^*(\epsilon^{\lambda T})|^2} \quad (54)$$

where it is understood that a weighted mean of $|c_{ss}(t)|^2$ is to be taken whenever $Re(\lambda)>0$.

Appendix II. Determination of the Mean-Square System Error for Periodic and Random Inputs

Case of a Sinusoidal Input

If the input is sinusoidal, say, $r(t) = \cos \omega t \equiv Re(\epsilon^{j\omega t})$, then, since the sampled-data system is linear, the steady-state output is

$$c_{ss}(t) = Re[K(j\omega;t)\epsilon^{j\omega t}] \quad (55)$$

where $K(j\omega;t)$ is the system function given by equation 50. Squaring this obtains

$$c_{ss}{}^2(t) = \{\{Re[K(j\omega;t)]\}\cos\omega t - \\ \{Im[K(j\omega;t)]\}\sin\omega t\}^2 \\ = Re^2[K(j\omega;t)]\cos^2\omega t + \\ Im^2[K(j\omega;t)]\sin^2\omega t - \\ Re[K(j\omega;t)]Im[K(j\omega;t)]\sin 2\omega t \quad (56)$$

Now the frequencies present in $K(j\omega;t)$ are direct current and multiples of the sampling frequency ω_0. Hence, if it is assumed that ω is not an exact multiple of ω_0 (which must be the case whenever the input and the sampler are independent), then the mean value of $Re^2[K(j\omega;t)]\cos^2\omega t$ is just $(1/2)\overline{Re^2}$

$[K(j\omega;t)]$, and the mean of the last term of equation 56 is zero. Thus the mean of equation 56 is

$$\overline{c_{ss}{}^2(t)} = \frac{1}{2}\overline{Re^2[K(j\omega;t)]} + \frac{1}{2}\overline{Im^2[K(j\omega;t)]} - 0$$
$$= \frac{1}{2}\overline{|K(j\omega;t)|^2} \quad (57)$$

Since the ripple factor is $(\overline{c_{ss}{}^2}/c_d{}^2)-1$, where c_d is the fundamental component of $K(j\omega;t)\epsilon^{j\omega t}$, it is seen that the ripple factor for a sinusoidal input is the same as for an input equal to $\epsilon^{j\omega t}$. Note that in both cases $\overline{|c_{ss}(t)|^2} = \overline{|K(j\omega;t)|^2} |r(t)|^2$.

If ω is an exact multiple of ω_0, then the system behaves just as if the input were direct current, the size of the output depending upon the phase of the input as well as its amplitude. However, such an instance will be rare, since the input frequency and the sampling rate will usually be independent. Note that the application of equation 53 to 57 gives, for a sinusoidal input of unit amplitude

$$\overline{c_{ss}{}^2(t)} = \frac{\sum_{n=-\infty}^{\infty} |G(j\omega+jn\omega_0)|^2}{2T^2|1+GH^*(\epsilon^{j\omega T})|^2}$$

Case of a Periodic Input

When the input is periodic, say, $r(t) = \sum_\mu a_\mu \cos(\omega_\mu t + \phi_\mu)$, then the output is

$$c_{ss}(t) = \sum_\mu a_\mu \{Re[K(j\omega_\mu;t)\epsilon^{j(\omega_\mu t+\phi_\mu)}]\} \quad (58)$$

where $\omega_\mu = \mu\omega_1$. Squaring this gives

$$c_{ss}{}^2(t) = \sum_\mu a_\mu{}^2\{Re[K(j\omega_\mu;t)\epsilon^{j(\omega_\mu t+\phi_\mu)}]\}^2 + \\ \sum_\mu\sum_{\substack{\nu \\ \mu\neq\nu}} a_\mu a_\nu \{Re[K(j\omega_\mu;t)\epsilon^{j(\omega_\mu t+\phi_\mu)}] \times \\ Re[K(j\omega_\nu;t)\epsilon^{j(\omega_\nu t+\phi_\nu)}]\} \quad (59)$$

Assuming that the difference of any integral multiples of ω_μ and ω_ν will not be exact multiples of the sampling frequency (which will be true whenever the input and the sampler are independent), then the second summation can be expressed as the sum of products of sinusoids of different frequencies, so that the mean of that summation therefore will be zero. The result obtained in the case of a sinusoidal input can be used to obtain the mean of the first summation of equation 59. Thus, for periodic inputs

$$\overline{c_{ss}{}^2(t)} = \frac{1}{2}\sum_\mu a_\mu{}^2 \overline{|K(j\omega_\mu;t)|^2} \quad (60)$$

This expression may be combined with equation 53 to obtain a more explicit equation.

It will be noted that equation 60 is also applicable to input functions composed of sinusoidal components whose frequencies are not harmonically related (i.e., functions which are "almost periodic"), so long as the input and the sampler are independent.

Case of a Random Input

If the input is a stationary random function (or even if its amplitude distribution

varies periodically with time) and if it is independent of the sampler, then a result of L. A. Zadeh's paper on correlation functions in variable networks[6] may be applied. Inserting the proper symbols in equation 23 of that paper gives

$$\psi_{cc}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{K(j\omega;t)K(-j\omega;t+\tau)} \times \\ S_{rr}(\omega)\epsilon^{-j\omega\tau}d\omega \quad (61)$$

where the bar indicates an averaging with respect to t, $S_{rr}(\omega)$ is the spectral density of the input, and $\psi_{cc}(\tau)$ is the correlation function of the output. The mean-square value of the output is obtained by setting $\tau=0$; thus

$$\overline{c_{ss}{}^2(t)} = \psi_{cc}(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{|K(j\omega;t)|^2} \times \\ S_{rr}(\omega)d\omega \quad (62)$$

Equation 40, which expresses the mean-square system error in terms of the spectral density of the input and the system's parameters, will now be derived. Taking the mean of the square of equation 1 gives, for the steady state

$$\overline{\epsilon^2(t)} = \overline{c_{ss}{}^2(t)} - 2\overline{c_{ss}(t)c_d(t)} + \overline{c_d{}^2(t)} \quad (63)$$

Using a procedure similar to that used in reference 6, equation 61 can be extended to cross correlation functions; in particular

$$\psi_{c_{ss}c_d}(\tau) \triangleq \overline{c_{ss}(t)c_d(t+\tau)} \\ = \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{K(j\omega;t)K_d(-j\omega)} S_{rr}(\omega)\epsilon^{-j\omega\tau}d\omega \\ = \text{constant} = \overline{c_{ss}(t)c_d(t)} \quad (64)$$

where

$$K_d(j\omega) \triangleq \frac{G(j\omega)}{T[1+GH^*(\epsilon^{j\omega T})]} \quad (65)$$

Using equations 50 and 65 it is seen that

$$\overline{K(j\omega;t)K_d(-j\omega)} = |K_d(j\omega)|^2 \quad (66)$$

Hence

$$\overline{c_{ss}(t)c_d(t)} = \frac{1}{2\pi}\int_{-\infty}^{\infty} |K_d(j\omega)|^2 S_{rr}(\omega)d\omega \\ = \overline{c_d{}^2(t)} \quad (67)$$

Combining equations 63 and 67 gives

$$\overline{\epsilon^2(t)} = \overline{c_{ss}{}^2(t)} - \overline{c_d{}^2(t)} \quad (68)$$

Applying equations 62 and 67 to 68 gives

$$\overline{\epsilon^2(t)} = \frac{1}{2\pi}\int_{-\infty}^{\infty} [\overline{|K(j\omega;t)|^2} - \\ |K_d(j\omega)|^2] S_{rr}(\omega)d\omega \quad (69)$$

By equations 50, 65, 32, and 34, one finally obtains

$$\overline{\epsilon^2(t)} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \times \\ \frac{\left[\sum_{n=-\infty}^{\infty} |G(j\omega+jn\omega_0)|^2\right] - |G(j\omega)|^2}{T^2|1+GH^*(\epsilon^{j\omega T})|^2} \times \\ S_{rr}(\omega)d\omega \quad (70)$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty} (\overline{e^2}j\omega)\, S_{rr}(\omega)\, d\omega \qquad (71)$$

which is the desired result.

References

1. THE ANALYSIS OF SAMPLED-DATA SYSTEMS J. R. Ragazzini, L. A. Zadeh. *AIEE Transactions* vol. 71, pt. II, Nov. 1952, pp. 225–34.

2. SAMPLED-DATA CONTROL SYSTEMS STUDIED THROUGH COMPARISON OF SAMPLING WITH AMPLITUDE MODULATION, W. K. Linvill. *Ibid.*, vol. 70, pt. II, 1951, pp. 1779–88.

3. EXTENSION OF CONVENTIONAL TECHNIQUES TO THE DESIGN OF SAMPLED-DATA SYSTEMS, W. K. Linvill, R. W. Sittler. *IRE Convention Record*, Institute of Radio Engineers, New York, N. Y., pt. 1, 1953, pp. 99–104.

4. FREQUENCY ANALYSIS OF SOME CLOSED-CYCLE SAMPLED-DATA CONTROL SYSTEMS, B. H. Stafford. *Report No. 3910*, Naval Research Laboratories, Washington, D. C., Jan. 21, 1952.

5. A LIST OF GENERALIZED LAPLACE TRANSFORMS, W. M. Stone. *Journal of Science*, Iowa State College, Ames, Iowa, vol. 22, Apr. 1948, pp. 215–25.

6. AUTOMATIC AND MANUAL CONTROL (book). Butterworth's Scientific Publications, London, England, 1952, "Characteristics of Sampling Servo Systems," C. Holt Smith, D. F. Lawden, A. E. Bailey, pp. 377–407.

7. FREQUENCY ANALYSIS OF VARIABLE NETWORKS, L. A. Zadeh. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 38, Mar. 1950, pp. 291–99.

8. CORRELATION FUNCTIONS AND POWER SPECTRA IN VARIABLE NETWORKS, L. A. Zadeh. *Ibid.*, Nov., pp. 1342–45.

An Adjustable Speed Power Selsyn System

S. Y. MERRITT
ASSOCIATE MEMBER AIEE

THIS paper treats a system consisting of two wound-rotor induction motors, with rotors paralleled through slip rings, the speed of which is controlled by a frequency changer connected to the slip rings. The system is a form of synchronous tie in which the driving motors and selsyns are incorporated in one machine. The block diagram in Fig. 1 gives a general picture of the system.

The slip frequency of the induction motors is the same as the output frequency of the frequency changer. An increase in this frequency results in a decrease in the speed of the induction motors, so that their speed can be controlled exactly by adjusting the output frequency of the frequency changer. The relation of speed to load, and speed control characteristics, are the same as those of two synchronous motors connected to an adjustable speed alternator.

This system would be useful in applications where two or more shafts must be driven at exactly proportional and adjustable speeds. Such applications are found in some paper, textile, and linoleum machinery, lift bridges, and kilns with associated feeders.[1]

Because of the possible uses, and the fact that this sytem has received very little treatment in the literature, development of an analysis and method of calculation is desirable. In this paper, means for calculation are developed and calculated results are compared with experimental results, expected performance is evaluated in a hypothetical application, and conclusions are drawn regarding system performance, design considerations, and need for further study. There are several power selsyn systems similar to the one being considered. Three of them will be discussed briefly.

With the frequency changer in the block diagram omitted, but the connection between the rotors remaining, the system represented is the one often called the synchronous tie. The induction motors tend to remain in the same angular position with respect to each other. The synchronous tie is usually used to synchronize shafts which supply power to meahanical loads and are driven by individual motors.[1,2]

In a second type of system, the frequency changer in the diagram is replaced by an impedance. With certain impedances, the induction motors can supply driving power as well as synchronizing power.[3-5]

In another application, the frequency changer may be replaced by an induction motor, often called a selsyn generator, having stator windings connected to rated voltage and rotor windings connected to the common rotor circuit of the other two induction motors. Motors so connected are often called selsyn motors. If the selsyn generator is driven mechanically, power is transmitted to the selsyn motors, and they are capable of driving and synchronizing loads at a speed proportional to that of the selsyn generator.[1] In this case the selsyn generator may be thought of as a kind of frequency changer.

The power selsyn system considered uses a frequency changer with an output frequency entirely independent of output current. An example of such a frequency changer is a rotor similar to that of a synchronous converter, having an iron ring pressed on it to complete the magnetic circuit, a polyphase voltage connected to the slip rings, and three brushes equally spaced around the commutator. If the rotor is driven externally, a polyphase voltage of frequency sf appears at the brushes, where s is the per-unit slip and f is the supply frequency. Frequency output is controlled by the speed of rotation. There is no interaction between the power required for rotation and the electric power flowing through the frequency changer. Therefore the mechanical power required by the frequency changer is just enough for friction and windage losses, and the output frequency is unaffected by changes in electric load on the frequency changer.

Operating Characteristics

1. The speed of both motors is equal, and can be adjusted by adjusting the output frequency of the frequency changer.

2. For a constant output frequency of the frequency changer, changes of load on the motors do not affect the speed. The developed torques must of course be within the steady-state stability limits, or the induction motors will lose synchronism with each other and the frequency changer. Curves of torque versus load angle can be plotted just as for synchronous machines to determine stability limits.

3. Motor power factor can be adjusted by adjusting the output voltage of the frequency changer. This has no effect on speed, within stability limits. The relationship of frequency changer voltage and motor power factor is the same as the relationship of field exciting voltage and power factor in synchronous motors. Over-all system power factor must be distinguished from motor power factor. The former is
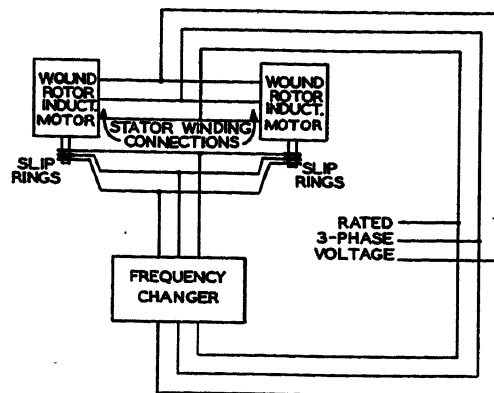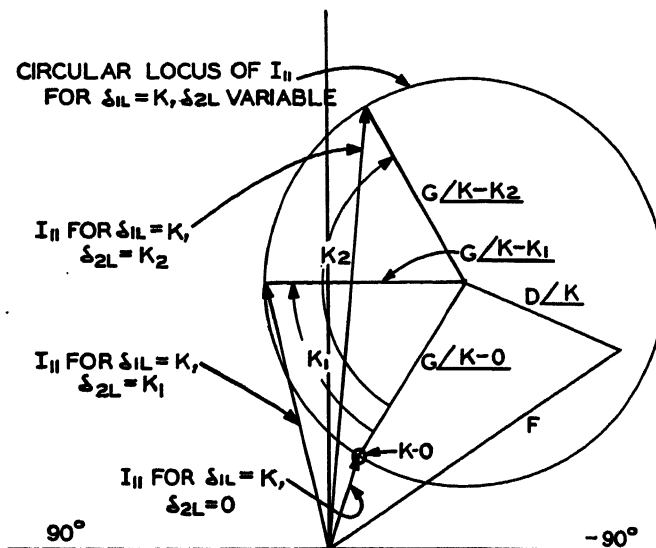
Fig. 1 (left).
Block diagram

Fig. 3 (right).
Illustration of stator current locus

influenced by the reactive power flowing to the frequency changer, which is generally opposite in sign to that flowing to the motors. For given conditions, an optimum condition of frequency changer voltage may be found, taking into account copper losses in motors and frequency changer.

4. Changes in the phase position of the frequency changer output voltage, such as a brush shift would produce, cause only a shift in the rotor positions of the motors. Speed, power input and output, complex primary current, and absolute secondary current of the motors are not affected after the transients have died out.

Mathematical Analysis

The circuit diagram is given in Fig. 2. The method of analysis, which is one used by Prof. P. C. Cromwell, is based on coupled circuits. The induction motors are assumed identical. All impedances are assumed to be balanced and constant. Stator and rotor impressed voltages are assumed to be balanced 3-phase positive sequence voltages. Core loss is neglected and a linear magnetization curve is assumed.

A physical concept can be expressed in terms of rotating field theory. Stator and rotor windings have the same number of phases and poles, and polyphase voltages of frequency $\omega/2\pi$ and $s\omega/2\pi$ respectively are applied to them, thereby inducing in the case of 2-pole motors magnetic fields revolving at ω and $s\omega$ radians per second respectively with respect to the windings which induce them. The fields "lock in" with each other, just as do the stator and rotor fields of synchronous machines. Thus with respect to a stationary reference point both stator and rotor fields have an angular velocity ω. The velocity of the rotor field with respect to the reference point is equal to the velocity of the rotor field with respect to the rotor plus the velocity of the rotor with respect to the reference point. Stated mathematically, this is $\omega = s\omega +$

$(\omega - s\omega)$. Thus the angular velocity of the rotor is $\omega - s\omega$. Though this approach is not related directly to the mathematical analysis, it furnishes an easily grasped physical concept.

Two basic rms current equations are derived: equation 1 is for the stator current and equation 2 is for the rotor current, both for machine 1

$$I_{11} = V_1/Z_1 + \frac{V_1 \cos(\delta_{1L} - \delta_{2L})/2}{(a^2 R_{T'}/s + R_1) + j(a^2 X_{T'} - X_1)}/2\alpha + (\delta_{1L} - \delta_{2L})/2 - \frac{jV_1 \sin(\delta_{1L} - \delta_{2L})/2}{(a^2 R_2/s + R_1) + j(a^2 X_2 - X_1)}/2\alpha + (\delta_{1L} - \delta_{2L})/2 - \frac{A^* V_2/\delta_{1L}}{s[(a^2 R_{T'}/s + R_1) + j(a^2 X_{T'} - X_1)]} \quad (1)$$

$$I_{21} = -A/\underline{-\delta_{1L}}(I_{11} - V_1/Z_1) \quad (2)$$

where

I_{11} = complex stator current of machine 1 (induction motor)

I_{21} = complex rotor current of machine 1 (induction motor)

δ_{1L} = angular displacement of rotor due to load, machine 1, from position occupied by rotors when $V_2 = 0$, and both rotors are loaded equally to given slip

δ_{2L} = angular displacement of rotor due to load, machine 2, as in the preceding item

s = per-unit slip

$A = a/\underline{-\alpha}$ = complex ratio of transformation of the induction motors. (a is the turns-ratio between stator and rotor windings. α is one-half the angle by which the axis of the exact induction motor circle diagram leads the 90-degree lagging position. α is usually small—1 or 2 degrees)

A^* = conjugate of A

R_1 = stator winding resistance per phase

R_2 = rotor winding resistance per phase

X_1 = stator winding inductive reactance per phase
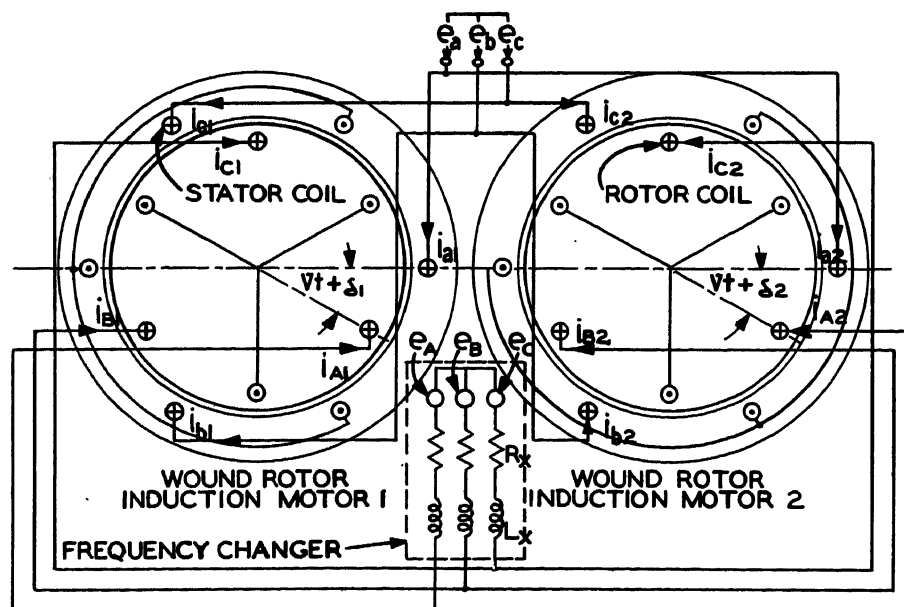
Fig. 2. Circuit diagram
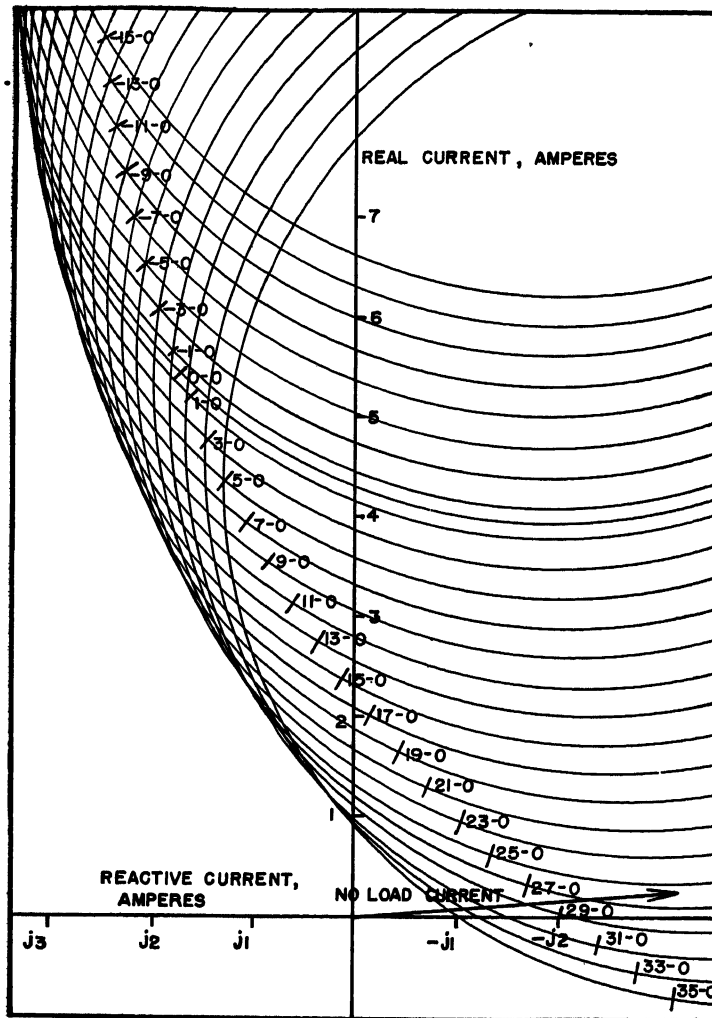
Fig. 4. Complete circle diagram for slip of 0.25

Fig. 5. Stator currents

Fig. 6. Power factors

X_2 = rotor winding inductive reactance per phase. (X_1 and X_2 are total reactances, not leakage reactances)

$(a^2X_2 - X_1)$ = leakage reactance per phase. (This is the sum of stator and rotor leakage reactances as used in the usual induction motor equivalent circuit, expressed in stator terms)

$Z_1 = R_1 + jX_1$ = stator winding impedance per phase. (This is the open-circuit impedance)

X_x = internal inductive reactance per phase of frequency changer

R_x = internal resistance per phase of frequency changer

$X_{T'} = X_2 + 2X_x$ and $R_{T'} = R_2 + 2R_x$. (These definitions are arbitrarily made for convenience in writing the equations)

V_1 = complex stator impressed voltage, rms line to neutral

V_2 = complex rotor impressed voltage, rms line to neutral

Equations for machine 2 may be obtained from those for machine 1 by interchanging δ_{1L} and δ_{2L} and changing the second current subscript to 2.

Without going into too much detail, a word of explanation is in order concerning $V_2 = v_2/\lambda$ and δ_{1L}, δ_{2L}. The mathematical development was in terms of absolute rotor angular positions δ_1 and δ_2. It was found convenient to choose as a reference point the position of the rotors with $V_2 = 0$ and the motors loaded equally to the given slip. Departures from this position due to load changes are defined as δ_{1L} and δ_{2L} and may be used directly in the equations. It should be noted that this convention results in some definite value of δ_{1L} and δ_{2L} for no-load conditions, as contrasted with zero value of the load angle for synchronous machines at no load.

Practically, λ is difficult to specify or measure since it is associated with a slip frequency voltage. Fortunately, it can be shown that λ has essentially no effect on the system operation. This can be proved mathematically, but is obvious, since a change in λ merely represents a momentary change in the output frequency of the frequency changer. It can be shown that if an effective value of $\lambda = h$ is calculated by the three steps to be described, and the real value of λ happens to be $h + g$, then the only error resulting is that the phase angle of the secondary currents is indicated as g radians too high. Correct values are indicated for all real and reactive power inputs and losses, complex primary currents, and absolute secondary currents. There is therefore no real need to know what λ is. The effective value of $\lambda = h$ is calculated as follows:
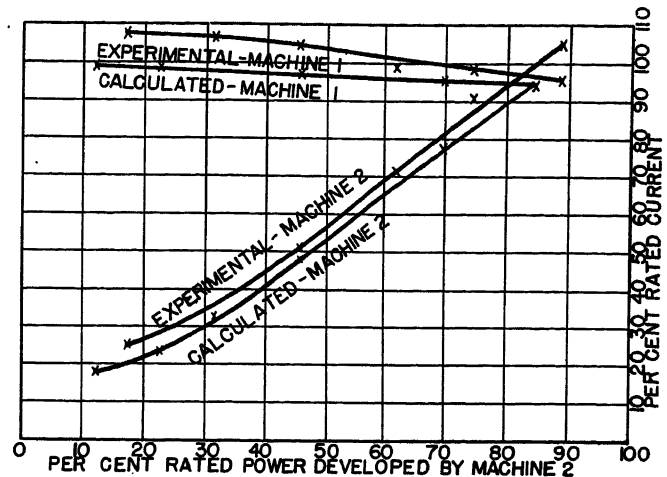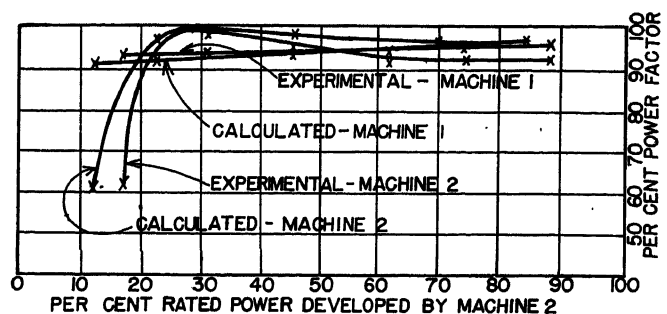
1. Assume $V_2 = 0$ and that both motors are loaded equally to the given slip s_h.

2. This is the reference condition, so that $\delta_1 = \delta_2 = \delta_{1L} = \delta_{2L} = 0$. If $V_2 = v_2/\lambda$ is now applied at frequency fs_h, the speed and power output will be unchanged.

3. The effective value h is whatever value of λ is required to give the same power output before and after the application of V_2. (The physical equivalent of these assumptions is merely that λ is adjusted before application of V_2 so that the frequency changer is synchronized with the existing secondary voltage before being cut in.)

Other conventions of λ, δ_{1L}, and δ_{2L} are no doubt possible, but this one gives good results and therefore has been used in the calculations for this paper.

The capital letters used on the right in equations 3 through 5 represent complex expressions which are constant at a given slip. For a given slip equation 1 may be written

$$I_{11} = E + B \cos (\delta_{1L} - \delta_{2L})/2$$
$$\underline{/(\delta_{1L} - \delta_{2L})/2} + C \sin (\delta_{1L} - \delta_{2L})/2$$
$$\underline{/(\delta_{1L} - \delta_{2L})/2} + D \underline{/\delta_{1L}} \quad (3)$$

Table I

	D-C System			Power Selsyn System		
Speed, rpm	600	900	1,200	600	900	1,200
System full load current, a-c amperes	54	73	91	54	68	103
System full-load power factor, per cent	100	100	100	87 lag	95 lag	93 lag
System full-load efficiency, per cent	61	67	72	71	75	68
Machinery required	two 28-kw 32-horsepower d-c shunt motors, one 56-kw d-c generator with pilot exciter, one 70-kva 63-horsepower synchronous motor			two 38-kva wound-rotor induction motors, one 75-kva frequency changer with fractional-horsepower adjustable speed driving motor		

which can be shown to be equal to

$$I_{11}=E+D/\delta_{1L}+jC+(B-jC)/2+$$
$$(B-jC)/2/(\delta_{1L}-\delta_{2L}) \quad (4)$$

or more simply

$$I_{11}=F+D/\delta_{1L}+G/(\delta_{1L}-\delta_{2L}) \quad (5)$$

For each value of δ_{1L} there is a circular locus of I_{11} as δ_{2L} is varied. Fig. 3 illustrates this, for $\delta_{1L}=k$ and all values of δ_{2L}, at some constant slip. Fig. 3 is not drawn to scale.

Fig. 4 is the complete circle diagram of I_{11} for δ_{1L} from 35 degrees to -15 degrees, and all values of δ_{2L}, for a slip of 25 per cent. The points -15-0, -13-0, -11-0. . .n-0, are similar to point k-0 in Fig. 3. Any point n-0 is located on the circular locus for $\delta_{1L}=n$. Given δ_{2L}, I_{11} can be found by stepping off the angle $-\delta_{2L}$ from n-0 around the circle as shown in Fig. 3. Fig. 4 applies to the experimental setup described in the following section, and was used in calculating its performance for Figs. 5 and 6.

Experimental and Calculated Results

The proposed system was set up and tested under load. Two similar 3-phase 6-pole wound-rotor induction motors rated at 5 kva, 220 volts, 60 cycles were used. A stator voltage of 117 was used for the induction motors because of the low output voltage of the frequency changer which was used. A frequency changer having the required characteristics was adapted from a 6-pole 6-kva Schrage brush shifting a-c motor.[6]

The system operated satisfactorily at 25-per-cent slip (900 rpm), at -10-per-cent slip (1,320 rpm), and at all intermediate speeds, including synchronous speed. Speed could be varied continuously between these limits by varying the output frequency of the frequency changer. Speed was not affected by load changes within the stable region. Speed could not be controlled by the frequency changer for speeds outside the range from 35-per-cent slip to -15-per-cent slip.

Undoubtedly this range would have been higher, and power developed without loss of synchronism would have been greater, if a frequency changer with a higher output voltage had been available. The one used had a maximum output voltage of about 30 per cent of the induction motor open-circuit secondary voltage.

The primary purpose of the experimental test was to serve as a check on the mathematical analysis. In Figs. 5 and 6, calculated and experimental values are compared for a slip of 25 per cent, developed power of machine 1 constant at 84.5 per cent of rated induction motor developed power, and developed power of machine 2 variable.

The following reasons for discrepancies between experimental and calculated curves are apparent:

1. About 50 per cent of the total loss was copper loss in the internal resistance of the frequency changer, which was definitely nonlinear because of brush drop.

2. There was some instrument error and some unbalance of impedances and voltages.

Although the errors may not have been unavoidable, it was not possible to eliminate them in the time available for the test. Making allowance for them, the experimental results indicate that the method of calculation is reasonable accurate.

Application

If a suitable frequency changer is available, a power selsyn system of this type can be designed for many applications normally requiring other types of drives. For illustrative purposes, this has been done for a hypothetical application and expected performance compared with that of an adjustable speed d-c system.

In the hypothetical application, which might be found in a cement mill, it is required to drive a kiln and kiln feeder at proportional adjustable speeds. Each load is normally driven at 900 rpm and requires approximately constant torque

at all speeds. The system must be capable of continuous operation at any speed between 600 rpm and 1,200 rpm and speed must be continuously variable under load between these limits. Comparison of expected performance of both systems at 600, 900, and 1,200 rpm with both loads equal is given in Table I.

D-c motors with adjustable voltage speed control are often used for kiln and feeder drives.[7] Wound-rotor induction driving motors, synchronized by a synchronous tie and having speed controlled by external resistance, are also used in this application. Because of the obviously greater efficiency of the d-c system at speeds substantially below synchronous the expected d-c system performance is considered, even though it does not have the exact synchronism feature of a power selsyn type of system. Expected performance of the d-c system is based on the assumption of a 440-volt 3-phase motor-generator set with a synchronous motor operating at 90-per-cent efficiency and 100-per-cent power factor and a d-c generator operating at 90-per-cent efficiency, supplying voltage adjustable between zero and rated value to two shunt motors operating at 88-per-cent efficiency. Efficiencies given are full-load values and are assumed to be reduced by 2 per cent at 900 rpm and by 5 per cent at 600 rpm.

Expected performance of the power selsyn system is based on the assumption of two identical 6-pole 3-phase 440-volt 60-cycle induction motors having a no-load current of $9.05/-85°$ $R_1=a^2R_2=0.300$ ohm per phase, leakage reactance of 2.3 ohms per phase, $A=2.0/-0.6°$, core loss of 600 watts, friction and windage loss of 500 watts at 1,200 rpm, 375 watts at 900 rpm, and 250 watts at 600 rpm. The frequency changer is assumed to have a magnetizing current of $6/-85°$, input voltage of 440 at 60 cycles, output voltage variable under load in steps from 55 to 220 at frequency variable from zero to 60 cycles, and internal impedance of $0.75+j0.106$ ohm per phase from the low-voltage side. Output frequency is assumed to be unaffected by load changes. Output voltage is assumed to be adjusted to 127 at 600 rpm, 69 at 900 rpm, and 59 at 1,200 rpm.

Conclusions

The circle diagram gives a convenient means for calculating performance. Comparison of calculated and experimental results indicates that the circle diagram is reasonably accurate. If a suitable frequency changer is available, a system can

be designed which will give good performance, with reasonable efficiencies and power factors over a wide range of speeds.

The whole system clearly depends on the existence of a frequency changer having a suitable voltage range and low enough internal resistance and reactance, which can be produced economically. Treatment of the design of such a machine is beyond the scope of this paper, but its possible uses have been demonstrated.

The equations developed could be modified to apply to other similar systems such as the modified Kramer drive used for wind tunnel drives[8] or to synchronous ties having impedance in the rotor circuit.

The most promising areas for further study of this system appear to be: 1 steady-state stability analysis; and 2: electrical and economic characteristics of the frequency changer.

References

1. SYNCHRONISM WITHOUT MECHANICAL CONNECTION, C. W. Drake. *Electrical World*, New York, N. Y., vol. 107, no. 8, Feb. 27, 1937, pp. 44–46.

2. INDUCTION MOTORS AS SELSYN DRIVES, L. M. Nowacki. *AIEE Transactions*, vol. 53, 1934, pp. 1721–26.

3. THE GEOMETRIC LOCI OF THE SYNCHRONOUS TIE, L. A. Finzi, L. C. Wellard. *AIEE Transactions*, vol. 68, pt. I, 1949, pp. 168–71.

4. THE TORQUES OF THE SYNCHRONOUS TIE—A STEADY-STATE ANALYSIS, L. A. Finzi, H. M. McConnell. *Ibid.*, pt. II, pp. 1147–52.

5. THEORY AND CALCULATION OF ELECTRICAL APPARATUS (book), C. P. Steinmetz. McGraw-Hill Book Company, New York, N. Y., 1917, pp. 159–65.

6. THEORY OF THE BRUSH-SHIFTING A-C MOTOR—I AND II, A. G. Conrad, F. C. Zweig, J. G. Clarke. *AIEE Transactions (Electrical Engineering)*, vol. 60, Aug. 1941, pp. 829–36.

7. STANDARD HANDBOOK FOR ELECTRICAL ENGINEERS, A. E. Knowlton, editor. McGraw-Hill Book Company, New York, N. Y., 1949, sec. 17–112, p. 1583.

8. A STUDY OF THE MODIFIED KRAMER OR ASYNCHRONOUS-SYNCHRONOUS CASCADE VARIABLE-SPEED DRIVE, M. M. Liwschitz, L. A. Kilgore, *AIEE Transactions (Electrical Engineering)*, vol. 61, May 1942, pp. 255–60.

Discussion

Thomas J. Higgins (University of Wisconsin, Madison, Wis.): I found this paper of considerable interest. The content is well organized; the purpose of the paper is clearly stated; the basic theory and pertinent approximation made in the course of development are well set out; and the account of the experimental investigation is nicely done.

Within the limitations stated by the author, the experimental data would appear to corroborate the basic analysis of the operation of the system studied.

P. C. Cromwell (The University of Tennessee, Knoxville, Tenn.): I congratulate the author in carrying through a rather complex problem in mathematics with the associated laboratory verification. The control system described is equivalent to a set of synchronous motors fed from one adjustable speed synchronous generator. It has two advantages over this system: 1. there are no starting difficulties since all units may be started and kept in space phase; and 2. the frequency changer must handle only the slip power and not the total power as a synchronous generator would be required to do. There would, however, be some limitation on the size of the installation due to the commutation of the frequency changer.

This paper is a good example of the use of symmetrical components in the analysis of a relatively complex problem in electric machinery. I have used this type of analysis in my classes, for both graduate and undergraduate students, for steady-state and transient problems for a number of years. It seems to me much sounder, from an educational standpoint, to start from Kirchhoff's laws and derive the equivalent circuit for a single-phase induction motor, e.g., to obtain its characteristic performance curves mathematically, than to use the usual physical concept approach favored by most textbooks.

It is my own opinion that the present disfavor of machinery courses has been brought about to some extent by the usual unanalytical approach now current in our colleges.

S. Y. Merritt: I wish to thank the discussers for their comments. Their interest is appreciated.

Electrical Grounding Systems and Corrosion

L. P. SCHAEFER
MEMBER AIEE

THE object of this paper is to consider the proper conduction of electric currents to the earth resulting in adequate grounding, and also conduction of currents through the earth resulting in corrosion or the prevention of corrosion. The functional objectives in designing an electrical grounding system are to provide a low resistance path for the fault currents to ground, and to provide safety for personnel by limiting the potentials to safe values. This is normal grounding practice and underground corrosion has generally been neglected until a considerable amount of damage has occurred to the primary structure.

The earth is tremendous in cross section and consequently when considered as a conductor between two electrodes it has a relatively low electrical resistance. However, an electrode placed into the earth will have a finite resistance value depending on the texture, aeration, moisture, and dissolved salts in the soil surrounding the electrode.[1,2]

Corrosion underground is an electrochemical reaction, often caused by dissimilar metals, dissimilar electrolytes, and stray currents. Dissimilar metal corrosion is set up as a result of two metals, such as copper and lead, which are electrically connected together and immersed in an electrolyte, such as a soil. One of the most common forms of this type of corrosion is that which occurs to the zinc case of a flashlight battery as the cell supplies $1^{1}/_{2}$ volts to an external load connected to the center carbon electrode and the zinc case. Corrosion resulting from dissimilar electrolytes is the type that occurs when soils of different resistivity surround the same section of buried metal pipe. Stray current corrosion occurs when stray electric currents leave a buried metal structure. Corrosion is halted when the corrosion currents which leave the structure are nullified or "bucked out" by externally applied current. The method of applying currents through the soil to halt corrosion on a structure is called cathodic protection. In order for the physical effects which are described in this paragraph to take place, it is necessary for electric

current to flow through the earth from one electrode to another. The earth is then a component of the corrosion circuit.

The earth as an electric circuit element can be measured in terms of its electrical resistivity. This value will vary with a number of factors, some of which were pointed out in the second paragraph. Also, the resistivity will vary with the depth below grade, where lower values are usually obtained as the soil becomes more moist because of the level of the water table. However, higher resistivities are also often experienced with increasing depth when rock structures are encountered. The electrical resistivity of the earth varies between extremely wide limits and is a decisive factor in using the earth as a return conductor such as for electrical grounding, corrosion, stray current electrolysis, inductive interference and lightning protection.

Earth resistivity measurements are made by the standard 4-pin or Wenner method.[3] This method uses four pins set up on grade where the two outside, or current, electrodes are 3D feet apart and the two center, or voltage, electrodes are D feet apart, where D is the distance between pins. The resulting values of voltage and current when used in the relation 191 $D E/I$ give the soil resistivity in ohms per cubic centimeter(ohm-cm). The particular spacing on the surface of the ground produces the average soil resistivity to the depth below grade corresponding to the surface spacing. The analysis for this method is given in Appendix I. Fig. 1 shows the variation of soil resistivity in ohm-cm, with depths from 2 to 12 feet below grade, for eight

widely separated parts of the United States.

The actual instrument used to measure the resistivity of the soil requires a manual or automatic means of reversing the polarity of the current, which is desirable because of polarization effects which occur at the electrodes. Commercial instruments indicate the ratio of voltage to a current. Usually, the depth of the electrodes in the soil is not important; however, if poor electrical contact is made with the earth, a resulting low instrument sensitivity will be obtained. Because of the wide variations in soil resistivity, it is almost always necessary to measure the resistivity in the precise locality under consideration. These same types of soil electrical resistivity tests can be used for geophysical prospecting, but their interpretation requires more detailed analysis than is necessary for the electrical grounding problem.

The electrode configuration which makes electrical contact with the soil should provide the lowest ohmic resistance which is feasible from an economic standpoint. This is true for electrical grounding systems, and is also a basic design criterion for ground beds which supply corrosion control for buried metallic structures. This concept, as shown later in the paper, makes it possible to combine good electrical grounding and adequate corrosion control.

Lead Cables in Ducts

The various types of underground lead cable sheaths which are exposed to external corrosion include electric power and

lighting cables, telephone, fire alarm, and teletalk cables. The usual construction is to pull the cables into fiber ducts imbedded in a concrete envelope. The most common construction provides separate duct runs and manholes for the various cable systems; however, some older systems may use common manholes and duct runs.

Corrosion of lead sheath cables is due to several causes, such as:

1. Small anodic and cathodic areas of lead due to splices, scratches, and abrasions.

2. Alternate wet and dry duct runs.

3. Galvanic effects due to copper bond straps, ground connections, or proximity to other metals.

4. Stray current effects from electrified railways, industrial chemical operations, and other sources.

In a number of known corrosive locations, it has been the practice in the past to use a protective neoprene sheath over the lead sheath to prevent sheath corrosion. This is a satisfactory mitigating measure against local cell action, dissimilar soil environment, and galvanic cells. The stray current problem, however, may require a complete system study resulting in the proper use of insulating splices and possible polarized forced drainage bonds. The neoprene sheath construction has proved satisfactory in providing a moisture-proof coating that apparently has a relatively small deterioration with age. It has been reported that polyethylene coatings also are most satisfactory.[4] One incident is known where termites have eaten through the polyvinyl chloride insulation on a buried cable in North Carolina, resulting in cable failure from moisture seepage.

The fundamental corrosion cell consists of two electrodes separated by an electrolyte on the internal circuit and electrically connected together on the external circuit. The coating practice described in the last paragraph supplies a high resistance electric barrier to current flow on the internal circuit, thus halting the corrosion. The insulated jacketed cable on the one hand requires adequate grounding because most certainly none will be obtained as a result of the underground cable run. The bare lead sheath cable on the other hand will have a fairly low ground resistance as a result of wet ducts and low soil resistivity. To this is added a grounding system to insure a low ground resistance value. The grounding system will generally consist of a bare copper ground bus of 500,000-circular-mil size or larger located in a spare duct or
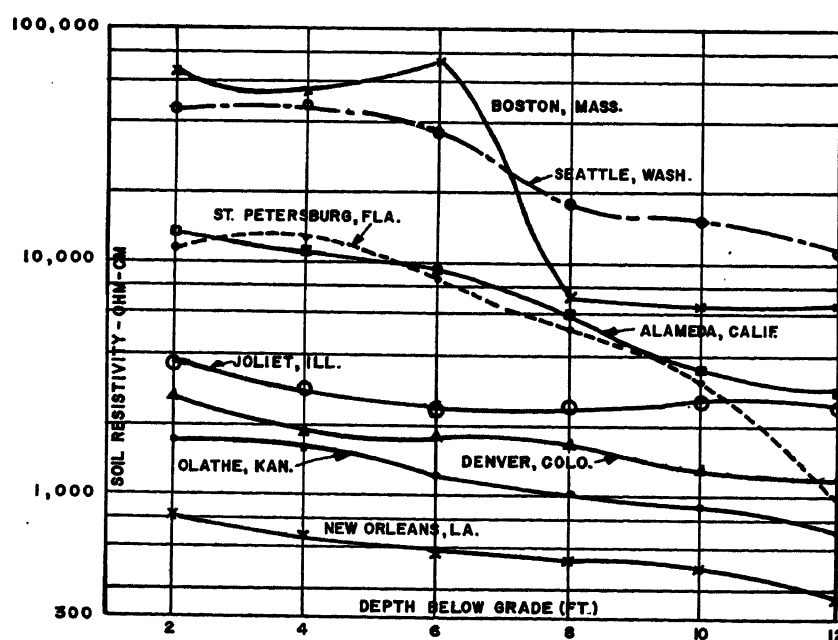
Fig. 1. Soil electrical resistivities at various depths to 12 feet for eight localities in the United States

embedded in the concrete envelope of the duct runs. Another usual practice is to locate terminal and switching equipment in underground vaults with a considerable ground mass, all of copper, located at the vault. Both the lead sheath and the copper ground rods and interconnecting cables are bare. A low resistance contact with the soil has thus been obtained for the grounding system. This is usually less than 5 ohms at the terminal equipment in the vault, and less than 25 ohms at each ordinary manhole. For the galvanic corrosion cell circuit, there are these low resistance contacts with the earth plus the driving potential of 0.5 to 0.6 volts from the lead-copper combination. The lead sacrifices to the copper, thus prolonging the life of the copper grounds at the expense of the lead cable sheaths.

A duct survey, made by placing a lead electrode in a spare duct, will show anodic sections of cable when voltages are measured between the cable and the lead electrode. These anodic sections will have the characteristic voltages given in Table I.

The stray current problems on lead sheath cables from electrified railways is rather complex. The track resistance to earth, the cable resistance to earth, soil resistivity, the geometry of the two systems, and rail bond continuity are all items affecting the magnitude of induced stray currents on the cables. The system study involves track-to-cable or track-to-soil potential recordings at all crossover locations and all areas where the track and cable runs are adjacent. On the basis of these recordings the locations of "natural" drainage bonds are determined, usually near the generating stations. Also, it is ascertained by the recordings whether the drainage bonds are adequate to supply corrosion protection or whether cathodic protection in the form of rectifier-driven "forced" drainage is to be added to the system of bonds connecting the cables and tracks.

Where power cable runs are bonded to a bare copper conductor in the duct envelope, or in a spare duct, there is certainly a problem in galvanic corrosion. Functionally, such a system may be very satisfactory from the standpoint of supplying a low resistance ground for operation of system relays. From a corrosion point of view, the very minimum of maintenance would require that the duct runs be pumped dry at all times. Where low soil resistivities prevail, it may be necessary to cut the lead cables free of the copper ground bus and apply new grounding using heavily zinc coated steel rods and cathodic protection. Such a system is de-

scribed in Appendix II. Where cable systems show a positive potential of 0.3 volt or more with respect to the environment, the condition is objectionable. If cathodic protection were attempted to be added to the lead cables which were thoroughly bonded to a copper ground bus or copper ground rods, it would be most difficult to change the potential of the cable with respect to the soil. This is because of the short-circuit path between the cathodic protection anodes and the many ground rods or ground bus, consequently making it difficult for cathodic protection current to reach the cables themselves.

The corrosion problems for underground telephone and fire alarm cables are approximately the same as those for power cables; however, the grounding is somewhat different. Telephone grounds are at exchanges and subscriber locations for operational purposes. The dissipation of lightning surges before there is damage to the cables depends on values and locations of grounds, spatial location of cable runs, and soil resistivity values of areas through which the cables run. Corrosion control on phone cables is somewhat simpler than for electric power cables because the grounding system is not so expensive and because the flashovers due to fault currents are absent. Underground fire alarm cables have much the same corrosion problems as do the telephone cables.

Substation Grounding

Electric substations are subject to external corrosion on noncurrent-carrying parts, such as housings, and above and below grade structures. The mechanical parts of air switch gear are subject to corrosion under conditions of moisture and weathering. This is also true of bus structures where dissimilar metals are involved.

Substation grounding systems are closely associated with corrosion problems. A low soil resistivity of 1,500 ohm-cm or less makes it relatively easy to

provide low resistance grounding; however, corrosion of the buried lead cables, column footings, steel conduit, etc., might be severe. From values of 15,000 to 150,000 ohm-cm, soil corrosion is reduced in severity, but it is most difficult to secure a low resistance ground without resorting to the use of an elaborate and expensive grid.

Major distribution substations usually require a ground resistance value of less than 1.0 ohm; minor substations are less than 5 ohms, and small capacity distribution equipment, such as transformers and switchgear, are grounded at less than 25 ohms. There are several available methods of measuring the ohmic value of substation grounds.[6] Geometrically speaking, the larger the grounding grid, the more elaborate is the test setup which is required to measure the resistance value.

Major substations will require low resistance grounding to provide for fast relaying to clear faults, and to provide for lightning and surge protection. To achieve these low values, a considerable quantity of bare ground rods and large quantities of bare connecting conductors are used. Such grounding systems are usually of copper, and as a result of this choice of material attention is then focused on the possible corrosion mitigating measures for the other underground and at grade structures. The ground non-current-carrying ferrous housings of transformers, switchgear, voltage regulators, etc., can usually be mounted so that no galvanic circuit prevails between the steel housings and the copper grounding system. Steel frame columns can be coated with bituminous paint and set in concrete to reduce corrosion possibilities. Care must be taken so that no salt is used with the concrete if poured in freezing weather. Underground cable and conduit systems should be brought into the substation in fiber ducts set in concrete. The duct runs should be kept pumped dry of water as part of a regular maintenance program. Underground cables and conduit should be grounded separately by use of heavily zinc coated steel rods, cathodically protected. The underground cable and conduit systems should also be coated with neoprene for the cables and a good bituminous paint system for the conduit where the substation is located in a low soil resistivity area. The coating provides an electrical resistance in the galvanic corrosion cell circuit which reduces the possible galvanic current which may flow. If the galvanic current is completely nullified, corrosion will not take place.

An electric substation using a con-

siderable amount of copper in a grounding system makes it extremely difficult to bring the other underground structures in the vicinity under cathodic protection. This is true because the copper ground system soil-to-structure potential must exceed 0.55 to 0.60 volt, as measured to a copper-copper sulfate reference half-cell, before any changes in the potential of steel structures to the soil can be obtained. This is related to the location of the two metals in the electromotive force series. Consequently, the steel and lead in the substation yard, especially if their surface areas are small compared to the copper grid, will receive little or no cathodic protection current while the copper grid system will be very much overprotected.

Attention then is brought to bear on possible alternative designs which might be used in place of the copper material where local conditions of corrosivity warrant the use of special design features. The major consideration for grounding conductors and rods is the ability to handle large fault currents up to 30,000 amperes for periods as long as 30 to 50 seconds without burning. Lightning surges with durations in the microsecond range are more of a mechanical problem than a heating one. Consequently, the material, other than copper, which might be used for a ground bed must have good current-carrying qualities and utilize fittings which are of the rigid clamp type. Usually the bare conductors connecting the ground rods contribute toward increasing the conductivity of the ground bed. Where the ground rods alone can supply the proper resistance. the connecting cables can be of copper, but these should be insulated with a neoprene sheath suitable for direct burial. Careful design attention should be given to the proper insulation of the connectors used to clamp the cable to the ground rods.

The remaining consideration is given to the proper choice of ground rod material. Zinc is used for ground rods, and is recommended in the 25-ohm range. Where lower resistance grounds are required to handle fault currents, a nonamorphous material is recommended. This choice is heavily zinc coated steel rods. The ground rods in turn are then cathodically protected by zinc anodes. Such a system will not only provide corrosion mitigation for the grounding system but will help protect the other buried structures and also will aid in reducing the total ohmic resistance level of the ground bed. Magnesium can be used as anode material in the place of zinc if a greater driving voltage is desirable.

Tower Line Grounding

A transmission line has a ground wire stretched between towers, with usually about 20 to 30 feet separation above the phase conductors at mid-span. Primary power distribution systems usually will be a grounded neutral system. The grounded neutral wire runs the entire length of the circuit and fans out with each single-phase lateral. The transmission line will usually have every tower grounded; this is true whether the towers are steel or wood H frames. Primary distribution is on wood poles, and ordinarily pole-butt grounds are used. The frequency with which grounds are located on the distribution lines will vary with the soil resistivity and insulation levels of the line. These locations will vary from an average of every fifth pole to every pole.

The problem of obtaining low resistance grounds at tower and pole locations is the same as that for the previous structures described in the paper. Rights of way and costs, which must be kept to reasonable values, are some of the problems in establishing low resistance grounds. In many locations where the subsoil or rock conditions prevent the use of long ground rods, it is necessary to use multiple short rods, buried horizontal wire, or a counterpoise system. Shallow grounds are subject to variations in resistance due to changes in temperature, such as frozen ground in the winter, or the moisture conditions of the soil. The tower-footing resistance for a transmission line will depend on the desired lightning performance of the line. A desired value of not over 10 ohms is not uncommon for tower-footing resistance. Primary distribution lines will have a desired pole resistance value of from 20 to 80 ohms, depending on the frequency of the grounding.[7]

Steel tower footings are subject to soil corrosion, dissimilar metal corrosion as a result of copper grounds being connected to the towers to secure low ohm resistance levels, long line currents as a result of dissimilar soil environments through which the transmission line footings are located, and stray current electrolysis as a result of

adjacent electrified railways and cathodic protection on pipe lines in the vicinity of the transmission lines.

Steel tower footings are subjected to considerable soil corrosion where they are located in soils with resistivities of less than 1,500 ohm-cm, and, as previously stated, such soils make it possible to supply a low resistance ground to the tower. Galvanic corrosion as a result of bonding dissimilar metals, usually zinc coated steel and copper, results in a rapid deterioration of the steel. Where the tower resistance value of, say, 10 ohms is obtained by copper ground rods, the zinc coating on the steel tower footings will act as an efficient sacrificial anode to protect the copper ground rods. As soon as the zinc coating is stripped, the steel tower footing will also start corroding. Where the transmission line leaves the substation, which contains a considerable amount of copper in the grounding grid, as many as 50 to 60 towers have been affected by the substation ground mass. This effect is measured by reading the change in tower-footing-to-soil potential, as measured with respect to a copper sulfate half-cell, before and after the tower ground wire is disconnected.

Long line currents, as a result of the dissimilar soil environments through which the transmission line passes, are produced in the grounding system of the transmission line. These direct currents enter the ground system at the towers located on high ground where the soil is less moist and has a higher electrical resistivity. The galvanic currents then leave the grounding system at the low points in the transmission line where the soil resistivity has a lower value. The mechanism of this type of corrosion is shown in Fig. 2, where galvanic currents flow to the line at towers A and from the line at tower B, resulting in tower-footing corrosion at B. A practical solution to the corrosion problem at tower B would be to revise the grounding to eliminate any dis-
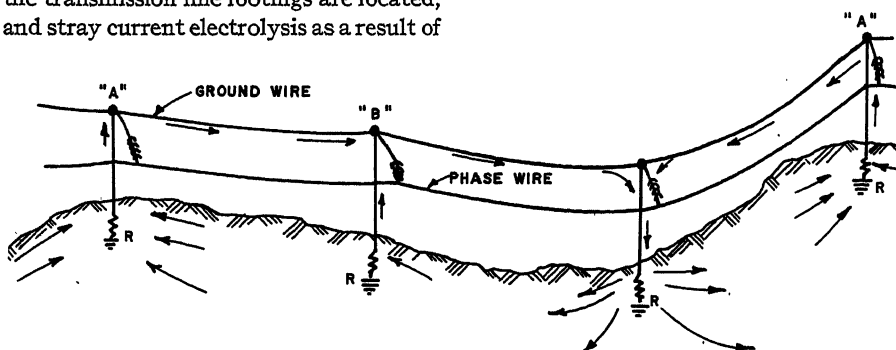

Fig. 2. Long line currents on grounding system of an electric transmission line

similar metal corrosion and then apply cathodic protection by using sacrificial anodes of zinc or magnesium. Approximately 5 milliamperes per square foot of tower and grounding surface, plus the additional current necessary to nullify or "buckout" the long line current, will be required of the sacrificial anode system. The current density of 5 milliamperes is based on the total area of the corroding metal and is a satisfactory value for steel in neutral soil; however, it is only the expected order of magnitude and the actual value can best be determined by field tests. The magnitude of the stray current flowing in the earth at tower B can be measured by an earth current meter, or by measuring the current-resistance drop along the ground wire on either side of the tower and then computing the current from the known resistance of the ground wire.

Stray direct currents superimposed on the transmission-line grounding system from sources such as electrified railways or as a result of cathodic protection installations on pipe lines, tanks farms, and other structures usually require system studies for a proper solution. Stray earth currents from any source, such as an electrified railway track or a cathodic protection ground bed, will return to the generating source by any adjacent low resistance path. The magnitude of such currents depends on a number of factors which are the same as those mentioned in the section entitled "Lead Cables in Ducts." The mitigating measures necessary to prevent stray current electrolysis of tower footings usually require consideration of a number of factors with due consideration being given to lightning and fault currents.

Power transmission lines can also be a source of damaging currents for buried pipe lines where insulating couplings are involved which are adjacent to, or which cross under, the transmission line. The pipe lines, by induction, become part of the transmission-line ground counterpoise system. Large currents which occur as a result of lightning flashovers and phase faults are induced on the buried pipe line, and the writer has measured as much as 80 amperes of alternating current induced on 6-inch piping running parallel to a 230-kv tower line for about 1 mile. Surge currents cause damage to the insulating flanges on the pipe lines by arcing across and fusing the flanges together.[8] Any kind of current on a pipe line transporting flammable fluids can be a hazard if the pipe line is cut or a section is removed and proper precautions are not taken.

Industrial Plant Grounding

Electrical grounding systems in industrial plants should maintain a low potential difference between equipment and machine frames, metal enclosures, building structures, etc., to avoid electric shock hazard to personnel, and to provide a low resistance path to earth to insure fast relay operation to clear fault currents before they become damaging. Electric currents as small as 0.05 to 0.17 ampere at 100 volts can be fatal to personnel. A good ground structure in an industrial plant will limit voltages across cable insulation, motor frames, and peak voltages in transient effects, and permits rapid isolation of faults.[9]

Grounding for industrial plants will follow the same general procedure as that for other installations. The best locations for the ground beds are determined by the soil electrical resistivity measurements, and these in turn supply design data necessary to define the ground beds. However, these better locations for the grounding system may not conform to the desired locations of substations or underground cable manholes. A major industrial operation will have a considerable amount of underground structure. This will consist of electric and telephone cables, air, water, fuel oil and sewer piping, underground and on-grade steel storage tanks, and steel building column footings. In a large industrial plant, there would be ample opportunity to use the underground piping for electrical grounding. From the standpoint of securing a good ground this might be satisfactory. Actually, pipe grounds are not good practice, since the piping may be abandoned or relocated or have nonconducting pipe joints, as is often the case with cast-iron water and sewer piping.

The use of nonferrous ground rods such as copper functionally will supply adequate grounding for an industrial plant. However, long corrosion-free life for the copper will be obtained at the expense of the underground steel and lead structures in a great many cases. On the other hand, the use of steel for ground rods presents a corrosion problem which obviously could not be solved by the use of some kind of coating on the steel ground rods. Consequently, cathodic protection in the form of sacrificial anodes of zinc or magnesium can be used to supply corrosion mitigation to the steel ground rods as well as the other underground structures in the vicinity. At the time an industrial plant is being constructed, it is relatively easy to install sacrificial anodes designed for a 20-year life, along with the required ground bed systems. A grounding system using cathodic protection of this type is shown in Fig. 3. If after a plant has been in operation for 5 or 10 years it is desired to revise the grounding system because of excessive corrosion repairs or other reasons, it may be more economical to use a rectifier supplied ground bed system and zinc coated steel ground rods.[10,11]

Most industrial plants will have some sort of stray current problem if no attention has been given to its control. The source of these stray direct currents may be welding equipment, plating processes, battery charging equipment, possibly some d-c control circuits, and motor generator equipment. The practice with welding equipment, which often causes trouble, is to ground one welding electrode to the building frame. This will superimpose the direct current onto the grounding system and, if the current which leaves the grounding rods is large enough, considerable corrosion will result. Such stray direct current will cause internal heating in the substation transformers if allowed to continue unchecked. Large program plating equipment, where a continuously moist atmosphere prevails, will have many leakage paths where the d-c bus structure is supported on insulators. Usually, since low voltages are involved, these insulators are most inadequately designed to prevent stray current
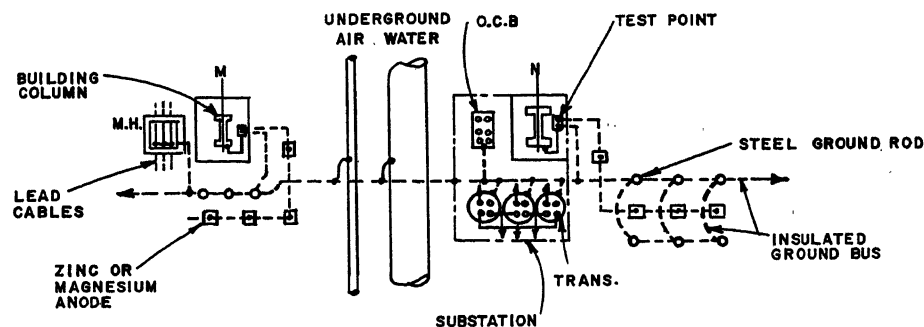

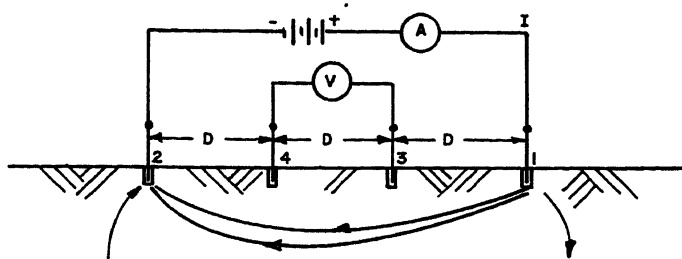Fig. 3. An industrial plant partial plan showing grounding and cathodic protection with sacrificial anodes

Fig. 4 (above). Four-electrode method of soil resistivity measurement

Fig. 5 (right). Cathodic protection on an underground lead cable system

leakage, as a result of moisture and dirt. Control circuits using direct current with one or the other polarity grounded, can readily superimpose direct current on the grounding circuits. Battery charging equipment is subject to d-c leakage, and the earth path can be of low resistance as a result of the available acid and acid fumes.

Lightning and Surge Protection

Direct lightning strokes are responsible for trouble on transmission lines. Grounded neutral systems have fairly low resistance paths to ground, and aerial telephone cables have lower failures from lightning because of the shielding effect of the sheath. If a lightning stroke has a crest value of 20,000 ampres, this current through a 5-ohm ground resistance produces a voltage drop of 100,000 volts, and through a 25-ohm ground this current value produces a 500,000-volt drop. About 90 per cent of this voltage will appear in a radius of 6 to 10 feet around the ground electrode.[12]

The actual peak values of current from lightning strokes are maintained for a short period of 20 to 30 microseconds. The surge current resulting from the lightning flashover producing a line-to-ground fault may last to 5 cycles until interrupted by the circuit breaker. Such fault currents require special precautions where cathodic protection systems are used in conjunction with electrical grounding systems.

Where sacrificial anodes are used to supply cathodic protection to the grounding system, no special fault current provisions are required since the anodes are an integral part of the grounding system. However, there are little or no data available on the impulse characteristics of commercial zinc or magnesium anodes, but it might be of considerable value to check the ground resistance of such anodes using a surge generator and the required measuring instruments.

The use of rectifier forced drainage systems in connection with grounding systems, or more usually for underground cable systems which are grounded, requires some special techniques. These devices have been described in the literature.[4,13] Buried pipe lines which become,

because of the geometry of the systems, either by inadvertant direct contact or by induction, part of the ground conductor system of a power transmission line will usually require spark-gap installations at the insulating flanges along the pipe line.

Static Grounding

The adequate discharge of static electricity by grounding is a problem from a hazard standpoint, where explosive and flammable materials are handled in connection with hospitals, fuel handling, ammunition depots, and many industrial processes. Hospital floors, which are conducting, usually require resistance values in the order of 100,000 ohms to 1 megohm between the floor and ground. Fuel handling grounds will be generally of the 25-ohm variety with the attendant switching equipment suitable for hazardous locations. Grounding systems for ammunition depots are rather elaborate for each individual manufacturing building, and their ohmic values and complexity are rather similar to those for electric substations.

Grounding systems in hospitals, due to their physical components, are generally not subject to corrosion. Fuel handling grounding practice consists of crossbonding the above ground piping and the connection of the ground bus to suitable grounding grids in order to prevent sparking in the presence of flammable vapors. For such systems the same types of corrosion problems will exist, in regard to dissimilar metals, cathodic protection, and stray currents, as were discussed in the previous sections of the paper.

Grounding systems at ammunition depots are designed to discharge electricity
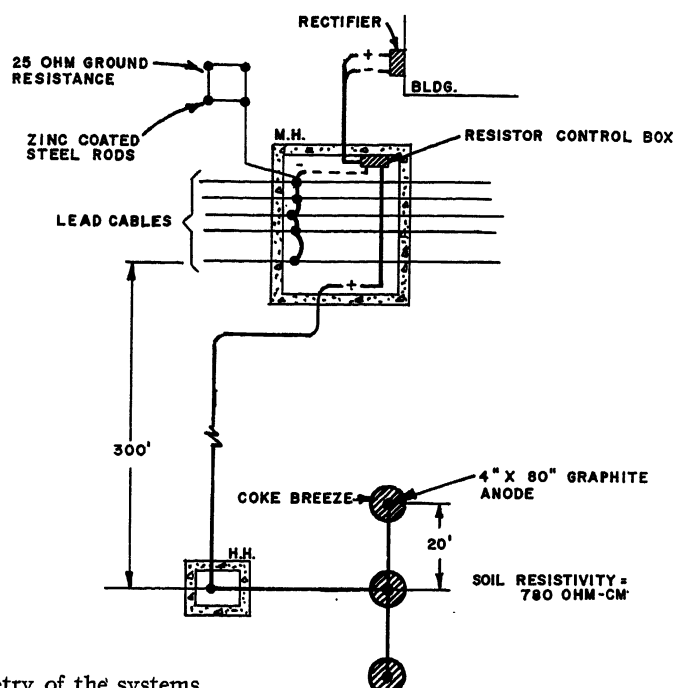
and to provide grounding for lightning. In addition to these precautions other safety practices are used, such as washing down walls to discharge charged particles and remove possible hazardous conditions. These are possibilities for corrosion in the systems and practices which may lead to malfunctioning and hence greatly reduce or nullify the safety features obtained by elaborate grounding systems. Each manufacturing building is generally isolated, because of the explosion hazard, and a grounding girdle of bare cable trenched 2 feet deep connects the various ground rods and the lightning rods. In a number of instances, these ground girdles are of bare 7/16-inch diameter stranded steel galvanized wire rope connecting to copper ground rods and steel lightning rod pole footings. Such dissimilar metal combinations are bad practice from the standpoint of the ground system itself. The copper ground rods also provide a chance for the corrosion of piping and lead cable systems which enter each building. In the design of such systems the use of dissimilar metals should be avoided and cathodic protection applied where corrosive conditions prevail.

Appendix I. Four-Electrode Method of Soil Resistivity Measurement[14,15]

In Fig. 4 let current I enter the ground at the point electrode no. 1 whose radius is s; area $= 2\pi s^2$; radial current density $J = I/2\pi s^2$; and let $\rho =$ soil resistivity. Then the electrical intensity $E(s)$ is

$$E(s) = \rho I/2\pi s^2$$

and the potential between the distance s and an infinite point is

$$V = \int_s^\infty E(s)ds = I\rho/2\pi s$$

The resistance of electrode no. 1 and an infinite point is

$$R(s) = V/I = \rho/2\pi s$$

The resistance of a circuit between 1 and 2 with a circuit between 3 and 4 is

$$R = R_{13} - R_{23} - (R_{14} - R_{24})$$
$$= \rho/2\pi(1/D - 1/2D - 1/2D + 1/D)$$
$$= \rho/2\pi D$$
$$\rho = 2\pi D R$$

where $R = V/I$.

For D in feet and $\rho =$ soil resistivity in ohm-cm

$$\rho = 2\pi 30.6 D \ V/I$$
$$= 191 D \ V/I$$

The depth to which the electrodes are placed into the earth will not affect the measurements, provided this distance is relatively small compared to the electrode spacing.

Appendix II. Cathodic Protection for Underground Lead Cable Systems

Sample calculation (see Fig. 5): A vertical anode buried in the soil has the following general equation for the resistance of the anode

$$r = \frac{0.0324\rho}{L} \text{ ohms} \tag{1}$$

where

$L =$ length of anode, feet
$\rho =$ specific resistivity of soil, ohm-cm
$r =$ resistance of the anode to remote ground
$N =$ anodes in a straight line at a separation S feet and connected in parallel

The combined resistance is

$$R = \frac{r}{KN}$$

K is 0.6 where $S = 20$ feet, K is 0.77 where $S = 40$ feet, K is 0.87 where $S = 60$ feet.

1. Basic data: Underground lead telephone cables are to be protected against corrosion by cathodic protection. Surface area of cables and ground bed involved: 5,640 square feet. The current density of 0.0003 ampere per square foot was determined by test to supply a change of 0.2-volt in the negative direction as measured from the cable to a lead shoe electrode in contact with the soil.
2. Total current required: $5,640 \times 0.0003 = 1.69$ amperes, or roughly 2 amperes.
3. The resistivity of the soil involved is 780 ohm-cm.
4. Impressed current system design:
(a) The anode resistance is computed using equation 1

$$r = \frac{0.0324}{L} \text{ ohms}$$

$$r = \frac{0.0324 \times 780}{8.7} = 2.89 \text{ ohms}$$

using 3–4-inch diameter by 80 inch long anodes with an effective length of 8.7 feet, as a result of the coke breeze backfill, and a separation of 20 feet. The ground bed resistance is

$$R = \frac{r}{KN} = \frac{2.89}{0.6 \times 3} = 1.6 \text{ ohms}$$

(b) Assuming a 3-per-cent voltage drop in the feed wire, a good general practice because of the relatively large amounts of current encountered, the voltage to be supplied by the rectifier is

$$E - 0.03 = IR$$
$$0.97E = 2 \times 1.6$$
$$E = 3.3 \text{ volts}$$

A rectifier with a d-c output of 5 or 6 volts and a rating of 10 amperes will be chosen for this application.

References

1. GROUNDING PRINCIPLES AND PRACTICE. I—FUNDAMENTAL CONSIDERATIONS ON GROUND CURRENTS, Reinhold Rudenberg. *Electrical Engineering*, vol. 64, Jan. 1945, pp. 1–13.

2. II—ESTABLISHING GROUNDS, Claude Jensen. *Ibid.*, Feb. 1945, pp. 68–74.

3. A METHOD OF MEASURING EARTH RESISTIVITY, Frank Wenner. *NBS Bulletin*, National Bureau of Standards, Washington, D. C., vol. 12, 1916, S258, p. 469.

4. CORROSION CONTROL OF UNDERGROUND POWER CABLES IN NEW YORK, F. E. Kulman. *AIEE Transactions*, vol. 73, pt. III, Aug. 1954, pp. 745–60.

5. DESIGN OF GROUNDING SYSTEMS TO MITIGATE CORROSION, TP-12 Committee. *National Association of Corrosion Engineers*, Houston, Tex., Mar. 11, 1953.

6. APPLICATION GUIDE ON METHODS OF SUBSTATION GROUNDING, AIEE Committee Report. *AIEE Transactions*, vol. 73, pt. III, Apr. 1954, pp. 271–77.

7. GROUNDING ELECTRIC CIRCUITS EFFECTIVELY. PART III—GROUND SYSTEM REQUIREMENTS, J. R. Eaton. *General Electric Review*, Schenectady, N. Y., vol. 44, Aug. 1941, pp. 451–56.

8. LIGHTNING ARRESTERS FOR THE PROTECTION OF INSULATED JOINTS IN BURIED PIPELINES, Stanley Kleinheksel. *Petroleum Engineer*, Dallas, Tex., vol. 22, Mar. 1950, pp. D87–D90.

9. EQUIPMENT GROUNDING FOR INDUSTRIAL PLANTS, L. J. Carpenter. *Electrical Engineering*, vol. 73, Mar. 1954, pp. 256–60.

10. GROUNDING PRINCIPLES AND PRACTICE APPLIED TO INDUSTRIAL PLANTS, E. B. Curdts. *Petroleum Industry Electrical News*, Feb. 1951. Reprinted for James G. Biddle Company, Philadelphia, Pa.

11. FIGHTING UNDERGROUND CORROSION AT THE NEW FAIRLESS WORKS, William E. Coleman, Harold G. Fostick. *Heating, Piping, and Air Conditioning*, Chicago, Ill., vol. 25, Apr. 1953, pp. 85–87.

12. GROUND CONNECTIONS FOR ELECTRICAL SYSTEMS, O. S. Peters. *NBS Technological Paper 108*, National Bureau of Standards, Washington, D. C., 1918.

13. GROUNDING AND CORROSION PROTECTION ON UNDERGROUND ELECTRIC POWER CABLE SHEATHS AND OIL- OR GAS-FILLED PIPE LINES, Robert J. Kuhn. *AIEE Transactions*, vol. 71, pt. III, Dec. 1952, pp. 990–93.

14. EARTH CONDUCTION EFFECTS IN TRANSMISSION SYSTEM (book), Erling D. Sunde. D. Van Nostrand Company, Inc., New York, N. Y., 1948.

15. SUBSURFACE EXPLORATION BY GEOPHYSICAL METHODS, E. R. Shepard. *ASTM Standard 96*, American Society for Testing Materials, Philadelphia, Pa., 1949.

Discussion

F. E. Kulman (Consolidated Edison Company of New York, Inc., New York, N. Y.): Mr. Schaefer's comprehensive paper discusses and clarifies the conflicting requirements of grounding and corrosion protection of underground structures. As a prelude to this discussion, it is well to restate briefly the reasons for grounding.

Grounding of electric equipment and circuits is necessary to prevent overvoltages caused by lightning, switching surges, and faults which would be harmful to personnel and damaging to equipment. Grounding also facilitates the quick operation of relays and circuit breakers to clear line-to-ground short circuits. Grounding wires for these purposes are designed to carry transient currents.

Where metallic pipes are exposed to the magnetic field of current-carrying conductors, it may be desirable to ground the pipe at both ends of the exposure to reduce the effect of induced voltages. Thus it is customary to ground the sheaths of telephone cables paralleling a high-voltage tower line to reduce the magnitude of currents (and noise) induced in the telephone circuits. Similarly, a buried coated pipe line paralleling an overhead electric circuit may be grounded at intervals to reduce the a-c potential difference between the pipe and the earth. Grounding wires for these purposes carry steady-state currents as well as transient currents.

Since grounding is necessary to protect human life, it receives and deserves precedence in electrical design. However, when the details of the grounding design are decided for any installation, they should be reviewed from the standpoint of hazards caused on buried structures, including the corrosion hazard. Mr. Schaefer has done a good job in pointing out these hazards and describing methods for mitigating them.

One of the subjects touched on is the problem of voltages induced in pipe lines paralleling high-voltage tower lines. Many of these pipe lines are provided with insulating joints for corrosion protection. It has been found that, under normal operating conditions on the electric line, a voltage in the order of 15 volts may be induced in relatively short sections of pipe line, less than 2 miles long.[1] This voltage is sufficient to give mild shocks to personnel operating valves on the pipe line. Larger voltages, of course, would be induced in longer sections of pipe line. Mitigative measures consist of reducing the length of pipe between insulating joints, and installing a local ground grid in the working area of the valve and connecting it to the main. In the case of a line-to-ground fault on the electric circuit

the induced voltages on the pipe may be much larger, possibly in the order of 1,000 volts or more.

Investigations made about 25 years ago[2] showed that the 60-cycle voltage induced in a telephone wire paralleling a faulted electric line is proportional to the ground return current and to the mutual impedance between the electric line and telephone wire. The mutual impedance depends primarily on the spacing between the two circuits and the electrical resistivity of the soil. For close spacings between the electric line and the telephone wire, the mutual impedance is in the order of 0.1 ohm per 1,000 feet of exposure. Thus, if the ground return current were 2,500 amperes, a potential difference of 250 volts could be induced in 1,000 feet of telephone wire or 1,320 volts per mile. Voltage of this magnitude can be expected in well-insulated pipe lines paralleling a high-voltage tower line.

Mr. Schaefer mentions that spark gaps have been installed across the insulating joints and that they have prevented burning of the insulated flanges by the fault current (see ref. 8 of the paper). Another method[3] consists of installing galvanic anodes along the coated pipe line to reduce the resistance between the pipe and earth, and grounding each side of the insulating joint to buried zinc electrodes. The zinc electrodes function as lightning arresters and also provide a low resistance path around the insulating joint for currents induced from the power line. Thus the method has the advantage of providing corrosion protection to the pipe line as well as grounding the pipe for lightning surges and induced voltages.

This discussion pertains to overhead electric tower lines only. In the case of underground pipe-type feeders where the pipe is solidly grounded at both ends, almost the entire fault current would be returned to the generating source by the pipe, and only a negligible portion would flow through the earth and other buried pipes.[4]

The problem of power cables supplying railway properties deserves special consideration with respect to grounding and corrosion. In this situation, precautions should be taken to keep the power cable sheaths and neutrals clear of contact with current-carrying parts of the railway and thus prevent the flow of stray current from the railway property to the power company cables. The railway properties referred to include railway substations, tracks, elevated structures, rapid-transit subways, and any structure connected with the railway system that might carry railway current.

Within railway substations it may be desirable to install an a-c ground bus to ground the neutral of low-voltage power services and the circuit-breaker housing, metering equipment, and transformer of high-voltage services. The a-c ground bus should be grounded to the water pipe but kept clear of contact with the substation building steel and the railway d-c and ground busses. Such construction would serve to reduce the electrolysis hazard on the power cables, and in the case of failure on the railway system reduce the flow of heavy fault currents which could cause burning of pipes, cables, and conduits.

REFERENCES

1. INSULATING JOINTS IN LONG PIPE LINES, PART 5: ELECTRICAL EFFECTS ON UNDERGROUND PIPE
LINES, J. P. Daratt. Gas, Los Angeles, Calif., vol. 30, no. 9, 1954, pp. 124, 126.

2. COUPLING FACTORS FOR GROUND RETURN CIRCUITS—GENERAL CONSIDERATIONS AND METHODS OF CALCULATION. Engineering Report 14, Joint Subcommittee on Development and Research of National Electric Light Association and Bell Telephone System, New York, N. Y., May 14, 1931.

3. GALVANIC ANODES CONTROL INDUCED VOLTAGES ON PIPE LINES, E. H. Thalmann. Corrosion, Houston, Tex., vol. 10, 1954, p. 367.

4. SINGLE-PHASE IMPEDANCE TO GROUND IN PIPE-TYPE CABLE, E. R. Thomas. AIEE Transactions, vol. 73, pt. III, Apr. 1954, pp. 336–44.

S. S. Watkins (Gibbs & Hill, Inc., New York, N. Y.): My comments are from the viewpoint of the engineer who is not a corrosion specialist nor a cathodic protection specialist but is concerned with the over-all design of substations and generating stations. Mr. Schaefer's paper renders a welcome assistance by warning that normal grounding systems can cause galvanic corrosion, by presenting useful data on corrosion, and by suggesting remedies. But I wish to caution against sacrifice of the objectives of the grounding system in the effort to prevent corrosion. There is an inherent conflict between the purposes of an effective grounding system and those of corrosion prevention.

The station designer uses copper cable and copper-clad rods for current capacity and permanence, designs an extensive system of ground rods and bare cable for low ground resistance and for limitation of voltage differences throughout the station area during ground fault, and welcomes any natural conditions of high water content, dissolved salts, and low resistivity of soil since these facilitate attainment of low ground resistance. But all of these factors can be causes of galvanic corrosion of underground steel and lead. The conflicting requirements of effective grounding and of corrosion prevention must be reconciled by the station designer.

Mr. Schaefer offers a solution in which all buried copper cables and buried copper connections of the grounding system are insulated and the ground rods are heavily zinc-coated and cathodically protected. No doubt this will prevent corrosion, but the insulation of the cable sacrifices an important function of buried bare cable, that of limiting the voltage differences throughout the station area during a fault to ground. Where buried cables are insulated, it may be necessary to increase the number of ground rods and space them closely along the station periphery and near mechanisms and switch handles to keep voltage differences within safe limits.

Published information on voltage gradients of actual grounding installations is scarce. Ground-resistance gradients usually can be measured, with standard ground-resistance instruments, during the period immediately preceding the energizing of a new station. The results of such measurements can be very valuable to the engineers concerned and valuable to the electrical industry as a whole if they are published.

Another solution of the conflicting requirements of corrosion prevention and effective grounding retains the equalizing effect of extensive bare conductors buried horizontally. Design calculations and field-test figures from a number of actual in-

stallations were given in a 1954 conference paper "All Steel Station Grounding Network" by S. J. Litrides.

Mr. Schaefer's mention of severe corrosion in soil of 1,500 ohm-cm or less resistivity suggests that high soil resistivity, which makes effective grounding difficult, may be a "cloud with a silver lining." It would be helpful to station designers if any figure of soil resistivity could be stated, above which bare copper station-grounding networks usually do not cause a serious degree of galvanic corrosion.

In the section on "Lightning and Surge Protection" it is stated that about 90 per cent of the voltage will appear in a radius of 6 to 10 feet around the ground electrode. I assume that this applies only to small electrodes, such as a single grounding rod.

The information which the paper gives on earth-resistivity measurements should prove useful. I agree that such measurements should be made at the precise location of every proposed grounding installation.

L. P. Schaefer: I wish to thank Mr. Kulman and Mr. Watkins for their constructive discussions. Both discussions have re-emphasized the reasons and the means of obtaining good electrical grounding which is of prime importance in the safe operation of electric equipment. It is well to point out here that performance is always a criterion of good electrical design and this should not be sacrificed because corrosion has caused failure somewhere down the line.

Mr. Kulman has done an excellent job of presenting data which are of value on the subject of induced voltages in buried pipe lines, cables, and pipe-type feeders. Fortunately, this was a subject of considerable interest some 20 years ago in the era of electrified interurban lines and street railways which are, as utilities, almost extinct at the present time. A buried pipe line is a leaky ground conductor which is subject to induced potentials where the pipe line parallels power transmission lines and electrified railway tracks. The induced voltage on the pipe line is a function of the current and the mutual impedance between the pipe and track or pipe and transmission line. The use of insulating joints on a pipe line which is in the field of a transmission line or an electric railway track system will depend on some of the following items:

1. Coating resistance of the pipe line.
2. Average soil resistivity.
3. Types of electric power transmission lines in the area.
4. Electric resistance to ground of tower footings, or track systems.
5. Geometric layout of pipe line and power transmission lines.
6. Location of insulating flanges.

Riordan[1] and Sunde[2] indicate the parameters and derive the general equations for solutions to the induced voltage problems for pipe lines from electric railway sources. These generalized equations are applicable to present-day cathodic protection problems on transmission pipe lines.

Mr. Watkins has pointed out the need for low potential differences between various pieces of electric equipment in a substation, and also the need for the reduction of volt-

age gradients to a minimum in and around any substation yard. For conditions of high soil resistivity it may be necessary to use an extensive underground counterpoise system of bare cable connecting driven rods or mesh mats in order to secure a low resistance ground. If the choice of material for this grounding system presents a possible galvanic corrosion cell problem, then it may be necessary to provide insulating sections in the underground lead cables or insulating joints in the pipe lines to break up the galvanic cell circuit. However, insulating sections may not be necessary if the mutual impedance of the grounding cable or grounding pipe system is high enough to prevent galvanic currents from flowing. Also, in low resistivity soils the total area of bare copper grounding may be small compared to the total area of buried lead cable. Con-

sequently, little or no damage will be done to the lead cable as a result of corrosion because of the small area of the copper cathode target in the galvanic cell circuit.

The copper of a bare copper connecting cable and copper-clad rods may be subject to environment corrosion. Cinders around the copper structure will provide a low resistance ground. However, the sulphuric acid will leak out of the cinders and cause corrosion of the copper. Also, as an example, copper can be subject to galvanic corrosion from cast-iron piping. It is a common practice to use copper service piping from cast-iron water mains. Unless an insulating coupling is used, the copper causes corrosion of the cast-iron pipe. This process removes the iron from the pipe, leaving a graphitized pipe near the copper service. The process is then reversed and the copper starts to

corrode to the graphitic pipe section.

Some of the problems described in the foregoing make it difficult to connect corrosion to a specific soil resistivity value. The soil resistivity is one component of the underground corrosion circuit, and, though important, there are other factors, such as the geometry of the electrodes, conditions of coatings, lineal resistance of the structures, and possible stray current problems which may also have a direct bearing on the corrosion process.

REFERENCES

1. CURRENT PROPAGATION IN ELECTRIC RAILWAY SYSTEMS, John Riordan. *AIEE Transactions*, vol. 51, Dec. 1932, pp. 1011–19.

2. CURRENTS AND POTENTIALS ALONG LEAKY GROUND-RETURN CONDUCTORS, E. D. Sunde. *Electrical Engineering*, vol. 55, Dec. 1936, pp. 1338–46.

An Analytical Method for the Design of Relay Servomechanisms

JOHN E. HART
ASSOCIATE MEMBER AIEE

RELAY servomechanisms in general show advantage over continuous servomechanisms with regard to cost, size, weight, and complexity. The application of relay servomechanisms, however, has been retarded by the meager understanding of obtainable operating characteristics and the lack of practical design methods.

The purpose of this paper is to present a method for design of relay servomechanisms. Physical circuitries and components are not considered directly, but, rather, the results are presented in nondimensional form to allow more general application. Certain assumptions have been made regarding the basic properties desired in the servomechanism to be designed. It is assumed desirable that the system be quiescent in steady state and that the inherent design simplicity be achieved. In order to execute rapidly the transient to a step input and minimize overshooting, it will be shown that a "critically damped" or "deadbeat" response is desired. (For a deadbeat response the relay will close and open only once when the system is subjected to an input step.)

In view of these assumptions, the method of design is based on derived analytical expressions of deadbeat criteria, wherein the parameters considered are restricted to the following: 1. inertia, 2. viscous damping, 3. motor torque,

4. gear ratio, 5. coulomb friction, 6. dynamic braking, 7. external braking, 8. relay sensitivity, 9. relay hysteresis, 10. relay transient time, 11. velocity feedback coefficient, and 12. velocity-squared feedback coefficient.

Nomenclature

DIMENSIONAL PARAMETERS

c = angular position of the motor shaft
J = system inertia reflected to the motor shaft
B = viscous friction coefficient reflected to the motor shaft
Q_m = stall torque of motor
T = motor speed time constant
$\dot{c}_m$ = maximum steady speed state of motor shaft in the absence of coulomb friction
a = pull-in voltage of relaying device, volts
h = error source sensitivity, volts per unit angle
N = gear ratio (ratio of motor speed to output speed)
k = linear velocity feedback coefficient, radians per radian per second
b = velocity-squared feedback coefficient, radians per radian per second²
T_d = relay drop-out time

NONDIMENSIONAL PARAMETERS

σ = angular position of the motor shaft; $c/(\dot{c}_m T)$
$\dot{\sigma}$ = angular velocity of the motor shaft; $\dot{c}/\dot{c}_m$
$\ddot{\sigma}$ = angular acceleration of the motor shaft; $\ddot{c}/(\dot{c}_m/T)$
ϵ = measure of one-half the width of the dead zone

λ = ratio of drop-out to pull-in voltage of the relaying device
κ = measure of linear velocity feedback coefficient
β = measure of velocity squared feedback coefficient
τ = ratio of relay drop-out time to motor-speed time constant
α = ratio of viscous friction coefficient when the relay is open to the viscous friction coefficient when the relay is closed
ϕ = ratio of coulomb friction present when the relay is open to the motor-stall torque
Δ = ratio of coulomb friction present when the relay is closed to the coulomb friction present when the relay is open

Deadbeat Criterion for Simplified Relay Servo

Consider a servo consisting of an error source, a polarized relay, and a motor, as shown in Fig. 1 The equation for the motor shaft is assumed to be

$$J\ddot{c} + B\dot{c} = Q_m f(c) \qquad (1)$$

where $f(c)$ is a function which is restricted to values of ± 1 or 0 and $\dot{c}$ and $\ddot{c}$ stand for dc/dt and d^2c/dt^2 respectively. The assumed characteristics of the polarized relay are shown in Fig. 2. Operation of the relay provides a discontinuous motor voltage, since the value of $f(c)$ becomes ± 1 whenever the magnitude of the error voltage exceeds the pull-in voltage of the relay. If the magnitude of the error voltage is less than the drop-out voltage
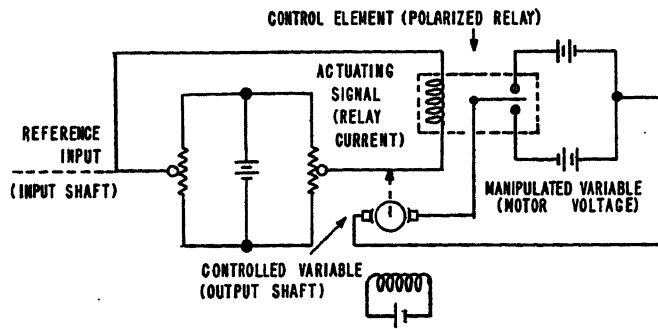
Fig. 1 (left). Simplified relay servomechanism

Fig. 2 (right). Relay characteristics
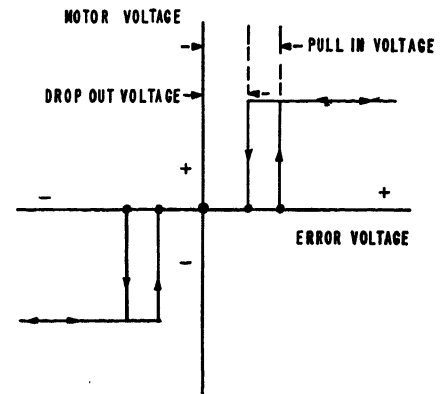
for the relay, the value of $f(c)$ is zero. Under the condition that the reference input remains zero, the error voltage will equal motor position multiplied by a proportionality factor involving error source sensitivity and gear ratio. The response of the system to a step error is therefore determined by assuming an initial value of the motor position different from the value of the reference input.

In order to simplify and nondimensionalize equation 1, divide through by Q_m. Then

$$\frac{\ddot{c}}{\dot{c}_m/T}+\frac{\dot{c}}{\dot{c}_m}=f(c) \tag{2}$$

where $T=J/B$ (the mechanical time constant) and $\dot{c}_m=Q_m/B$ (the maximum steady-state speed of the motor). Equation 2 may be rewritten as

$$\ddot{\sigma}+\dot{\sigma}=\pm 1 \text{ or } 0 \tag{3}$$

Use will also be made of a nondimensional measure of shaft position σ, where σ is $c/\dot{c}_m T$.

The response of the system is described by curves of the successive states of position and velocity in the phase plane. Such curves are termed trajectories. Fig. 3 shows a phase plane where the boundaries of relay operation are delineated and the equations for these boundaries, as well as the equations of motion of the system in various regions, are indicated. In the equations describing the boundaries for relay operation, ϵ is a nondimensional measure of the relay pull-in voltage

$$\epsilon=\frac{aN}{h\dot{c}_m T}$$

For the system under consideration, the equations for the trajectories in the various regions are given by Weiss.[1] In the region where the relay is open, the equation is

$$\dot{\sigma}+\sigma=\dot{\sigma}_0+\sigma_0 \tag{4}$$

where $\dot{\sigma}_0$ and σ_0 are the initial coordinates. In Fig. 4, if the system is to respond to all step errors regardless of magnitude without reversal of the

motor, the velocity must come to zero when the position is less than ϵ. From nondimensional considerations, the initial velocity, in the central region, can have a maximum value of unity. The initial value of position for the region where the relay is open is $-\lambda\epsilon$. Inserting these conditions in equation 4 gives the inequality

$$0+\epsilon>1-\lambda\epsilon \tag{5}$$

or

$$\epsilon>\frac{1}{1+\lambda} \tag{6}$$

This inequality is therefore the criterion for deadbeat response. Obviously, overshooting may occur, and an error may exist in steady state; but the magnitude of overshoot and the value of steady-state error are restricted to a range corresponding to plus or minus the pull-in voltage of the relay.

Practical Systems

In practical applications it may be desirable to reduce the dead zone necessary for deadbeat operation below the value required for the simplified relay servo. Such a reduction may be provided by altering the boundary equations for relay operation or by altering the equations of motion for the various states of the relay. Alteration of the boundary equations may be accomplished by velocity feedback or lead networks; the equations of motion may be affected by employing various types of braking or utilizing mechanical damping means.

If a resistance-capacitance lead network is employed, the gain of the system is adversely affected. Consistent with the assumed properties of simplicity, light weight, etc., for the complete system, the requirement of additional amplification associated with a resistance-capacitance network renders this type of stabilization undesirable. In many cases mechanical damps, such as the Lanchester damp, are also undesirable, because of the resulting limitation on

acceleration. The maximum motor speed is not affected by the presence of a mechanical damp but the transient time is substantially increased.

Two forms of velocity feedback have been suggested in the literature.[2] The simpler of these consists of subtracting, from the error signal, a signal proportional to the rate of change of error or the velocity of the output. In the presence of this type of feedback the equation for the relay operation boundary of Fig. 4, e.g., would change from

$$\sigma=-\lambda\epsilon \tag{7}$$

to

$$\sigma=-\lambda\epsilon-\kappa\dot{\sigma} \tag{8}$$

where

$$\kappa=\frac{k}{T}$$

The more complex form of velocity feedback consists of subtracting from the error signal a signal proportional to the square of the rate of change of error or the square of the velocity of the output. For this case the relay boundary would be, e.g.

$$\sigma=-\lambda\epsilon-\beta\dot{\sigma}^2 \tag{9}$$

Consistent with the nondimensional form

$$\beta=\frac{b\dot{c}_m}{T}$$

It is evident that, when the relay is open, the presence of a braking action will tend to stabilize the system. In equation 1 it was assumed that the value of viscous damping B was independent of the operation of the relay. In reality, when the relay is open, the value of viscous damping depends upon the amount of dynamic braking employed. The nondimensional equation applicable when the relay is open will be

$$\ddot{\sigma}+\alpha\dot{\sigma}=0 \tag{10}$$

The response of the system will be affected by the value of α. The system
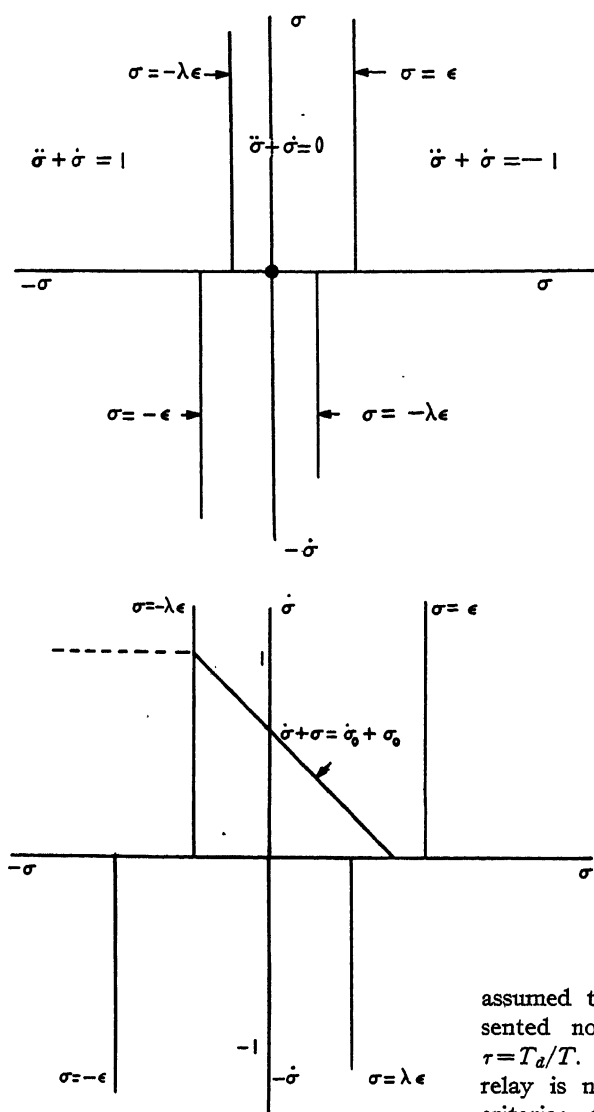
Fig. 3 (left). Equations of motion and boundaries for regions of phase plane

Fig. 4 (left, below). Deadbeat response and trajectory equation

Fig. 5 (right). Limits of deadbeat operation
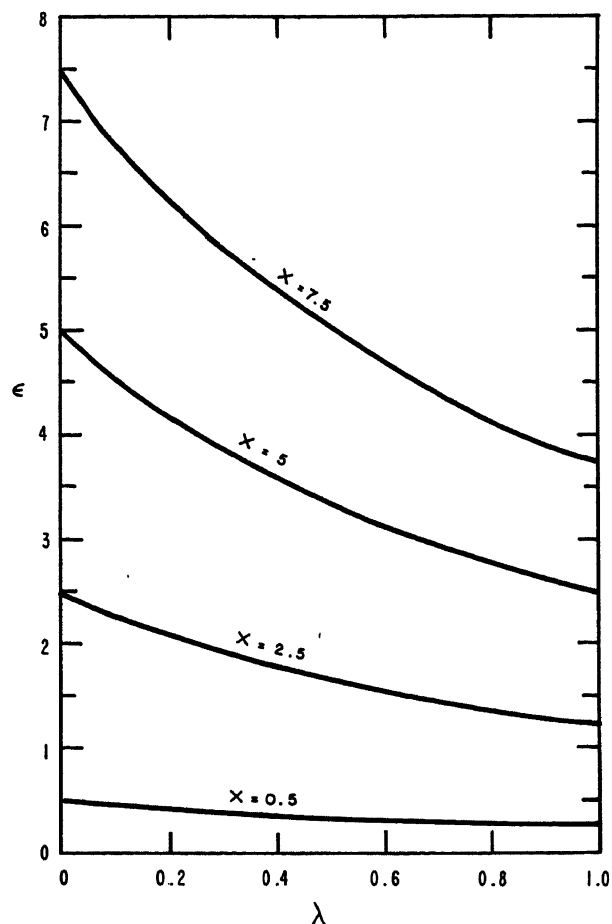
will be unstable if α is equal to zero, unless some other type of braking action is employed. Such supplementary braking action may be obtained by applying a friction brake to the motor shaft when the relay is open. Assuming the frictional force of the brake to be a constant (and the ratio of this torque to the motor torque to be ϕ) the equation of motion when the relay is open will be

$$\ddot{\sigma} + \alpha \dot{\sigma} = -\phi \qquad (11)$$

Deadbeat Criteria

The method of derivation described for a simplified relay servo has been employed to establish the more comprehensive criteria for deadbeat operation; see Appendix I. One final parameter is considererd, namely, the drop-out time of the relay. The drop-out time is defined as the time required for the relay contacts to open after the coil voltage reduces to a value below the drop-out voltage. The drop-out time T_d is

assumed to be constant and is represented nondimensionally as τ, where $\tau = T_d/T$. The pull-in time for the relay is not involved in the deadbeat criteria; the value of this parameter does influence the response of the system, however, as will be discussed in the next section.

The deadbeat criterion when simple velocity feedback is employed is

$$\epsilon > \frac{\tau + \dfrac{1}{\alpha} - \kappa - \dfrac{\phi}{\alpha^2} \ln\left(1 + \dfrac{\alpha}{\phi}\right)}{1 + \lambda} \qquad (12)$$

For velocity squared feedback the criterion is

$$\epsilon > \frac{\tau + \dfrac{1}{\alpha} - \beta - \dfrac{\phi}{\alpha^2} \ln\left(1 + \dfrac{\alpha}{\phi}\right)}{1 + \lambda} \qquad (13)$$

If α is equal to zero the criteria are

$$\epsilon > \frac{\tau - \kappa + \dfrac{1}{2\phi}}{1 + \lambda} \qquad (14)$$

and

$$\epsilon > \frac{\tau - \beta + \dfrac{1}{2\phi}}{1 + \lambda} \qquad (15)$$

Graphical presentation of these criteria

may be simplified by considering th general criterion

$$\epsilon > \frac{x}{1 + \lambda} \qquad (16)$$

where x is the numerator of the right-hand side of the inequalities 12 through 15. Curves for the inequality 16 are shown in Fig. 5 for several values of x. The value of the numerator x as a function of $(\tau - \kappa)$ or $(\tau - \beta)$ for various values of ϕ is shown in Fig. 6 for $\alpha = 0$ and in Fig. 7 for $\alpha = 1$. Sets of curves for other values of α are easily determined.

Physical Interpretations of Nondimensional Parameters and Assumptions

The criteria for deadbeat response presented are exact in so far as the equations utilized in their derivation apply to the physical system under consideration. In reality, however, a physical system may not be described by simple linear equations and the factors neglected may, under certain circumstances, be significant in limiting the performance that may be obtained. Other factors, which do not affect the deadbeat criteria, must also be con-
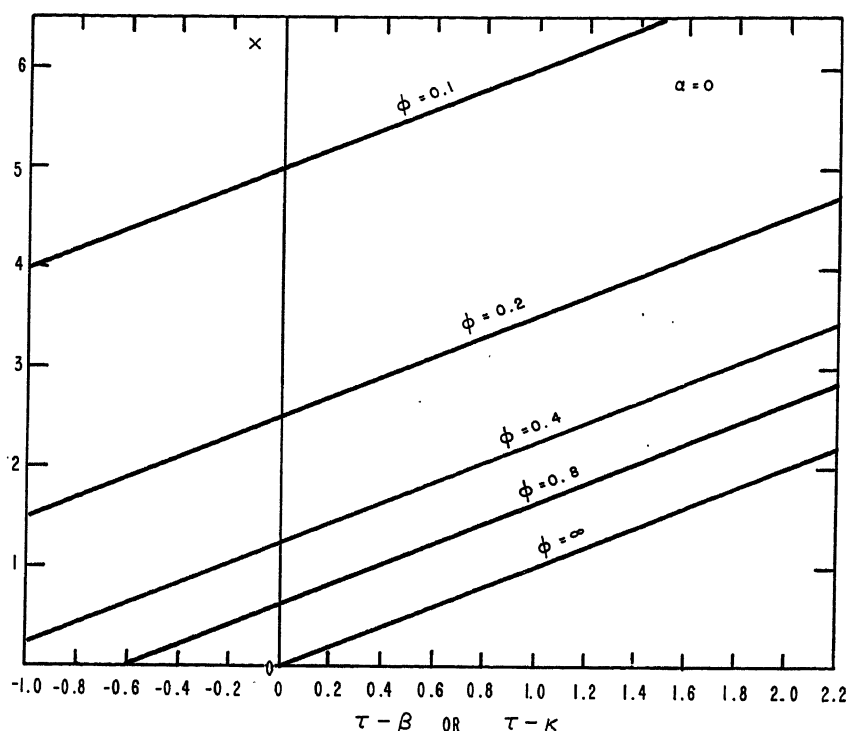
Fig. 6. Deadbeat factor with no dynamic braking

sidered because of their influence on the response time.

The relation of the physical parameters to a particular value of a nondimensional parameter may be determined by inspection. If the expression for ϵ is reduced to the fundamental parameters, e.g.

$$\epsilon = \frac{aNB^2}{hQ_mJ}$$

it is seen that ϵ increases proportionally to the relay pull-in voltage a. If a is increased, in order to increase ϵ, the actual dead zone at the output shaft increases. If, on the other hand, ϵ is increased by increasing the viscous damping B, the actual dead zone will be unchanged; the time constant will be decreased, however, as will the maximum motor speed. Decreasing inertia J increases ϵ and shortens the response time, but produces no effect on the maximum motor speed.

An understanding of these and other relations is necessary to provide an insight to the fundamental properties of relay servomechanisms. If, e.g., J is decreased, the resulting increase in ϵ is favorable; the decrease in J increases τ, however, and thereby adversely affects the system response.

It has been assumed that the time required for the motor torque to build and decay is negligible. In a physical system this time constant may not be neglected if a rapid response and mini-

mum dead zone are to be obtained. From a phase-plane consideration, the presence of this time constant limits the rate of change of velocity with respect to position at the transition points of the trajectory. However, the motor-torque time constant may be approximated as a time delay and added to the relay transient time.

If a system is deadbeat, the presence of a pull-in time for the relay or the build-up time of motor torque will not alter the deadbeat characteristic. The time required for the system to respond to an input step, however, will be increased by the presence of such time constants or time delays.

Two conditions will result in a steady-state oscillation or limit-cycle operation of a relay servomechanism, notwithstanding its compliance with a deadbeat criterion. If a load torque greater than ϕ is applied to the output shaft, the relay will operate intermittently so that the time average of motor torque plus ϕ will cancel the applied torque. In like manner, when the input is a constant velocity, the relay will operate intermittently to provide an average output velocity equal to the input velocity. The actual deviation between the input and output in these cases obviously depends upon the system parameters. In general, it may be stated that the smaller the actual dead zone and the higher the motor torque, the smaller the actual deviation will be.

In a physical system, it will generally be found that a design requiring the actual dead zone to approach zero, or the ratio of drop-out to pull-in voltage to approach unity, will result in a limit-cycle oscillation. One reason for such oscillation is the statistical nature of the relay operation. E.g., the saturation of an electromechanical relay, because of
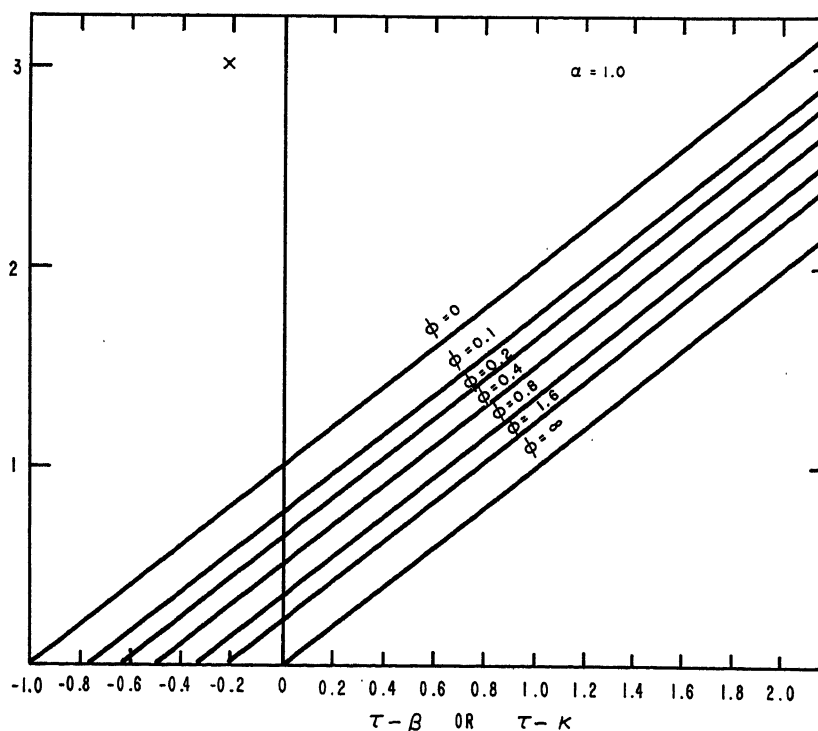
Fig. 7. Deadbeat factor with 100-per-cent dynamic braking
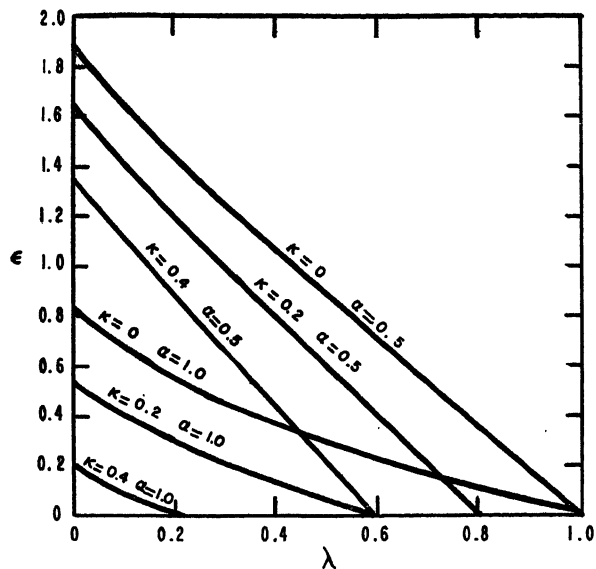
Fig. 8 (left). Stability limits with linear velocity feedback
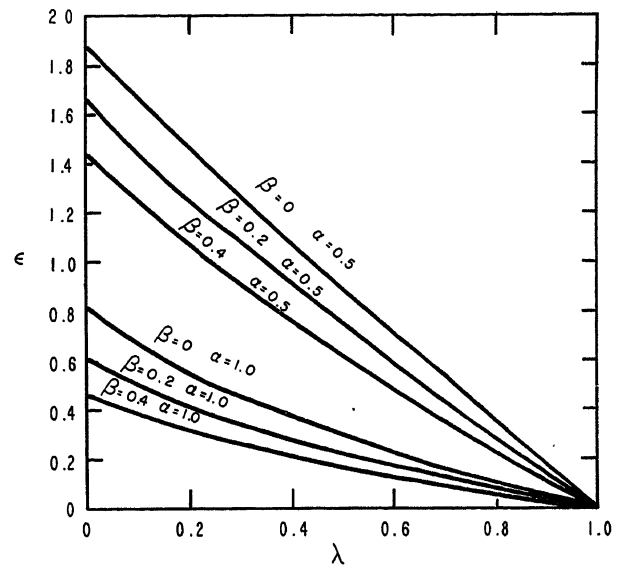
Fig. 9 (right). Stability limits with velocity-squared feedback

the momentary application of a high voltage, can cause a variation in the relay operation.

Linear and squared-velocity feedback, as stabilizing means, have been mathematically analyzed by Feldbaum.[3] The results given by Feldbaum have been expanded and are presented in Figs. 8 and 9. In these figures, points above or to the right of a curve indicate sets of parameters corresponding to stable systems. Relay drop-out time may be included in stability considerations but the parameter ϕ leads to simultaneous transcendental equations.

In Figs. 6 and 7 the use of the feedback factors κ and β, as to their effect on deadbeat operation, is indicated. From the abscissa axes of Figs. 6 and 7 it is apparent that the parameters κ or β may be utilized to nullify the effect of relay drop-out time τ. If $(\tau - \kappa)$ or $(\tau - \beta)$ is less than zero, the system will be further improved. The extent to which these factors may be employed to improve the system is limited, however, by secondary considerations. Consider the case of velocity squared feedback as shown in Fig. 10, where the value of β is relatively large. For this case, the trajectory of the response following a step input indicates that the relay will open and close several times—but in the same sense. Such operation is considered undesirable because of the general increase in response time involved. Inspection of Fig. 11 indicates that this type of operation may also be encountered with linear velocity feedback. A cursory analysis shows, however, that the value of κ which may be employed before this "overdamped" type of response is encountered is greater than the value of β. This factor, together with a consideration

of the circuit complexity required to provide a squared term, renders the linear velocity feedback method more desirable in most cases. In support of the use of the squared term, however, it should be noted that its employment results in a significant decrease in transient time when the major portion of the braking action is of the coulomb friction type.

One additional factor should be mentioned before presenting the design procedure: the coulomb friction present which is independent of the operation of the relay. Such friction may arise from gear loading or other sources inherent in the system. In designing, an attempt is made to minimize this friction because of its adverse effect on motor torque. Expressions are developed in Appendix I for deadbeat criteria where coulomb friction is not negligible.

Design Procedure

The specifications to be met by a servomechanism may be stated in various ways. For linear systems, the form of the specifications is of little importance because of the implicit relation of one form to another. For relay servomechanisms, however, it is desirable that the specifications state the allowable static and steady-state following errors, the allowable overshoot, and the maximum transient times. Specifications of bandwidth or other frequency characteristics would require a complex interpretation because of the nonsinusoidal form of the output of a relay servomechanism when subjected to a sinusoidal forcing function and also because of the effect of input magnitude on the output wave shape.

The design of a relay servomechanism is carried out by first calculating the motor power required from considerations of the maximum motor speed, the load characteristics, and the approximate accelerations required. After selecting a motor, the values of c_m and T may be calculated from the motor and load parameters. The sensitivity required for the error source and relay pull-in voltage may be calculated to give the specified value of dead zone. From these data the value of ϵ may be calculated. With the selection of a relay, or pilot and power relays, the value of λ is obtained. Assuming the system to be deadbeat, the required value of x is obtained from the inequality 16. The value of τ is determined for the relay assumed and this parameter, together with the required value of x, permits determination of various sets of the remaining parameters: α, ϕ, and κ or β from curves on Figs. 6 and 7. This information and knowledge of the relay pull-in time will indicate alternate arrangements which may be utilized to meet the required specifications. As in any servomechanism design, testing of a mockup of the system will be necessary for adjustment and to indicate allowable values of backlash, resolution requirements, and other secondary considerations.

This design method was employed in calculating the values of ϵ for deadbeat operation in two relay servomechanisms. The values of the various nondimensional parameters and the measured and calculated values of ϵ for these systems are presented in Table I. The approximate agreement achieved for these servomechanisms indicates the usefulness of this design method. The system referred to as servo no. 1 in Table I could be
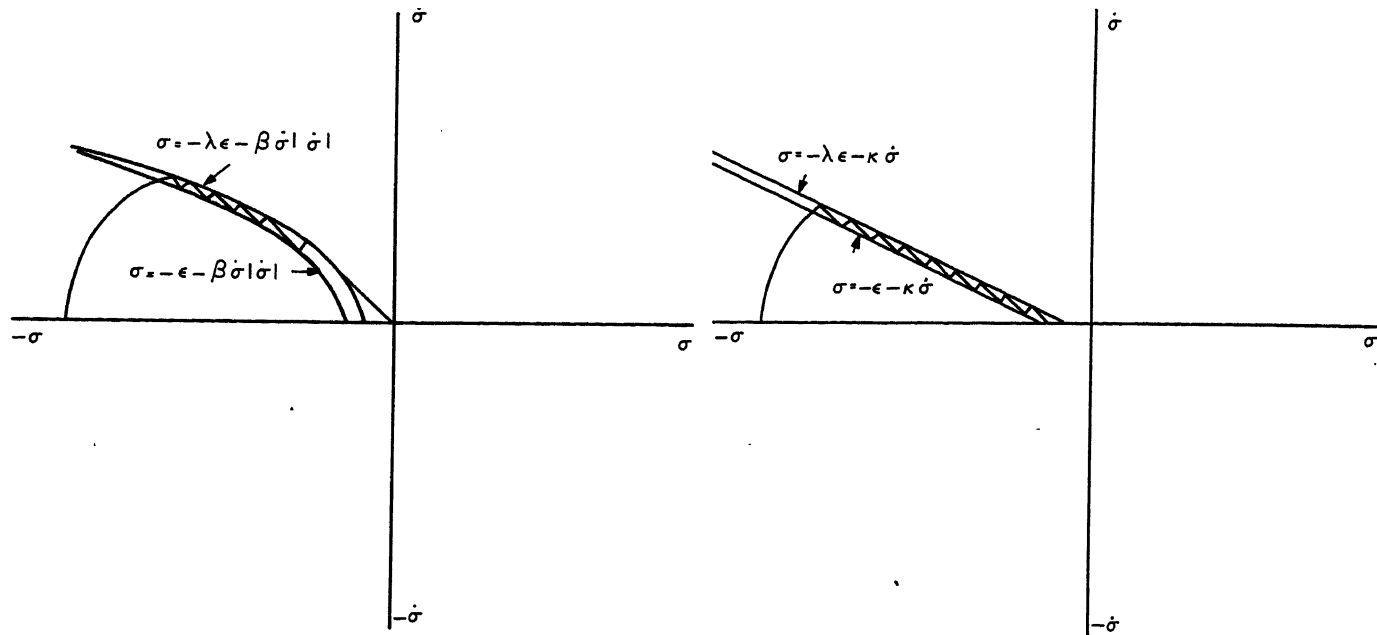
Fig. 10. Typical response with excessive velocity-squared feedback   **Fig. 11.** Typical response with excessive linear velocity feedback

adjusted to meet the following specifications:

Static error: ± 15 minutes of arc.
Maximum slewing speed: 150 degrees per second.
Velocity error coefficients: 0.05 at 10 degrees per second; 0.0035 at 150 degrees per second.
Transient time for 90-degree step input: less than 1 second.
Maximum overshoot for 90-degree step input: 30 minutes of arc.

In general, it is found that a dead zone equal to about half that required for

deadbeat operation may be employed. This procedure results in a peak overshoot magnitude approximately equal to the total width of the dead zone.

Summary

The deadbeat criteria presented form the basis of a method of design of relay servomechanisms. The complexity of the systems considered has been limited to include only the essential parameters because of the assumption that a relay

servomechanism would not be employed except to utilize its inherent property of simplicity. The presentation of criteria in terms of relations of non-dimensional parameters permits application of this design method to a wide range of physical systems.

For systems defined by the parameters considered, the general criteria represent the necessary and sufficient conditions for deadbeat operation. In attempting to obtain general criteria for a more complex system, which considers the

Table I. Calculated and Measured Parameters for Physical Models of Two Relay Servomechanisms

Parameter	Servo No. 1	Servo No. 2
B, foot-pounds per radian per second	8.39×10^{-5}	8.39×10^{-5}
J, slug feet^2	0.284×10^{-5}	6.76×10^{-5}
Qm, foot-pounds	3.0×10^{-2}	3.0×10^{-2}
T, seconds	0.03385	0.806
a/h, radians	0.025	0.0476
N, gear ratio	60 to 1	210 to 1
k, radians per radian per second	0.0238	0.1588
Coulomb friction, foot-pounds	1.07×10^{-2}	1.393×10^{-2}
Braking friction, foot-pounds	1.07×10^{-2}	1.393×10^{-2}
Dynamic braking, foot-pounds per radian per second	1.2×10^{-5}	1.2×10^{-5}
Relay pull-in voltage	3.5	3.5
Relay drop-out voltage	2.1	2.1
Relay drop-out time (approximate), seconds	1.0×10^{-3}	1.0×10^{-3}
α	0.143	0.143
κ	0.703	0.197
ϕ	0.358	0.465
τ	0.03 (approximate)	0.0012 (approximate)
λ	0.6	0.6
Δ	1.0	1.0
ϵ calculated	0.0313	0.0681
ϵ measured	0.124	0.0346

$$\dot{\sigma}_1 + \alpha \sigma_1 - \frac{\phi}{\alpha} \ln(\dot{\sigma}_1 + \frac{\phi}{\alpha}) = \sigma_2 + \alpha \sigma_2 - \frac{\phi}{\alpha} \ln(\dot{\sigma}_2 + \frac{\phi}{\alpha})$$

$$\sigma = -\lambda \epsilon + (1 - \Delta\phi)[(1 - \epsilon^T)(1 - \kappa) + \tau] + [\epsilon^T(1 - \kappa) - 1]\dot{\sigma}$$
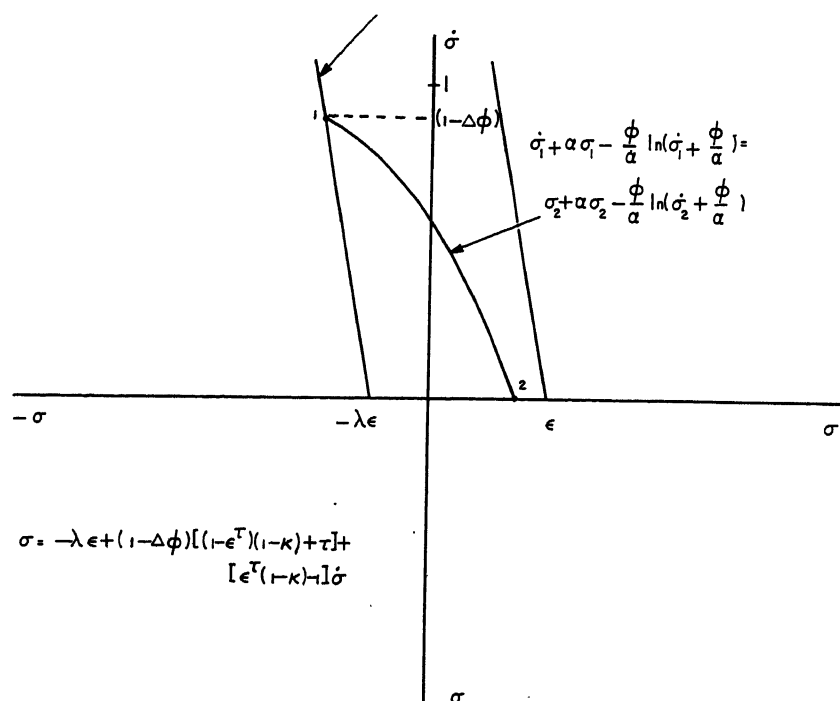
Fig. 12. Typical deadbeat response with maximum initial velocity

motor-torque time constant or the effects of a lead network, e.g., it is found that the solution of simultaneous transcendental equations is required. Thus the designer of the more complex relay servos will be required to use empirical or approximate methods of design.

Appendix I. Derivation of Deadbeat Criteria

Deadbeat criteria will be derived for systems with dynamic braking α, coulomb braking ϕ, coulomb friction Δ, relay drop-out time τ, and either linear velocity feedback κ or velocity-squared feedback β.

For these systems the equation of motion where the relay is open is

$$\ddot{\sigma} + \alpha\dot{\sigma} = -\phi \qquad (17)$$

The corresponding equation for the trajectories in this region is

$$\dot{\sigma}_1 + \alpha\sigma_1 - \frac{\phi}{\alpha}\ln\left(\dot{\sigma}_1 + \frac{\phi}{\alpha}\right)$$
$$= \dot{\sigma}_2 + \alpha\sigma_2 - \frac{\phi}{\alpha}\ln\left(\dot{\sigma}_2 + \frac{\phi}{\alpha}\right) \qquad (18)$$

For the case where $\alpha = 0$ the equation of motion will be

$$\ddot{\sigma} = -\phi \qquad (19)$$

and the corresponding equation for the trajectories

$$\sigma_1 + \frac{\dot{\sigma}_1{}^2}{2\phi} = \sigma_2 + \frac{\dot{\sigma}_2{}^2}{2\phi} \qquad (20)$$

The relay boundary equation is used in the derivation of each of the deadbeat criteria. The form of this equation, as influenced by the drop-out time τ, will be indicated before the consideration of several criteria.

Weiss[1] shows that if the equation for the locus of points on the phase plane corresponding to a given event is

$$\sigma_1 = m + n\dot{\sigma}_1 \qquad (21)$$

then the equation for the locus of events on each trajectory, following the initial event by a constant time delay τ, is

$$\sigma = m + D[(1-e^\tau)(n+1)+\tau] + \\ [e^\tau(n+1)-1]\dot{\sigma} \qquad (22)$$

where D is the right-hand side of the non-dimensional equation of motion for that region.

For the system with linear velocity feedback, the boundary equation in the absence of a time delay is (for point 1)

$$\sigma_1 = -\lambda\epsilon - \kappa\dot{\sigma}_1 \qquad (23)$$

The equation of motion for the trajectories approaching this boundary is

$$\ddot{\sigma} + \dot{\sigma} = 1 - \Delta\phi \qquad (24)$$

The value of D in equation 22 is therefore $1 - \Delta\phi$. Substituting this value for D and the values of m and n from equation 23 into equation 22 gives the boundary equation for point 1 as

$$\sigma_1 = -\lambda\epsilon + (1-\Delta\phi)[(1-e^\tau)(1-\kappa)+\tau] + \\ [e^\tau(1-\kappa)-1]\dot{\sigma}_1 \qquad (25)$$

If a system is deadbeat for the case where the velocity (upon entering the region where the relay is open) has its maximum possible value, then it will be deadbeat for all step inputs. The value of the maximum velocity is

$$\dot{\sigma}_1 = 1 - \Delta\phi \qquad (26)$$

A phase plane showing the boundary and trajectory equations and a typical deadbeat response for this case are shown in Fig. 12. For the system to be deadbeat it is necessary that σ_2 be less than ϵ when $\dot{\sigma}_2$ is zero. Substituting equations 25, 26, $\dot{\sigma}_2 = 0$, and $\sigma_2 < \epsilon$ into equation 18 gives, after simplification, the desired criterion

$$\left(\tau + \frac{1}{\alpha} - \kappa\right)(1-\Delta\phi) - \frac{\phi}{\alpha^2}\times \\ \epsilon > \frac{\ln\left[1 + \frac{\alpha(1-\Delta\phi)}{\phi}\right]}{1+\lambda} \qquad (27)$$

Similarly, these substitutions in equation 20

give the deadbeat criterion for the case where $\alpha = 0$

$$\epsilon > \frac{(1-\Delta\phi)\left(\tau + \frac{1}{2\phi} - \kappa - \frac{\Delta}{2}\right)}{1+\lambda} \qquad (28)$$

For a system with velocity-squared feedback the boundary equation for point 1, in the absence of relay drop-out time, is

$$\sigma_1 = -\lambda\epsilon - \beta\dot{\sigma}_1{}^2 \qquad (29)$$

In the presence of drop-out time (from equation 22), the boundary equation is

$$\sigma_1 = -\lambda\epsilon + (1-\Delta\phi)[(1-e^\tau)(1-\beta\dot{\sigma}_1)+\tau] + \\ [e^\tau(1-\beta\dot{\sigma}_1)-1]\dot{\sigma}_1 \qquad (30)$$

Substituting equation 30 and the conditions $\dot{\sigma}_1 = 1 - \Delta\phi$, $\dot{\sigma}_2 = 0$, and $\sigma_2 < \epsilon$ into equation 18 gives, after simplification, the deadbeat criterion

$$\epsilon > \frac{\left[\tau + \frac{1}{\alpha} - \beta(1-\Delta\phi)\right](1-\Delta\phi) - \frac{\phi}{\alpha^2}\ln\left[1 + \frac{\alpha(1-\Delta\phi)}{\phi}\right]}{1+\lambda} \qquad (31)$$

Similarly, these substitutions in equation 20 give the deadbeat criterion for the case where $\alpha = 0$

$$\epsilon > \frac{(1-\Delta\phi)\left(\tau + \frac{1}{2\phi} + \beta\Delta\phi - \beta - \frac{\Delta}{2}\right)}{1+\lambda} \qquad (32)$$

References

1. ANALYSIS OF RELAY SERVOMECHANISMS, H. K. Weiss. *Journal of Aeronautical Sciences*, New York, N. Y., vol. 13, no. 7, July 1946, pp. 364–76.

2. AUTOMATIC AND MANUAL CONTROL (book). Butterworth's Scientific Publications, London, England, 1952, "Stabilization of On-Off Controlled Servomechanisms," A. M. Uttley, P. H. Hammond, pp. 285–99.

3. THE SIMPLEST RELAY SYSTEM OF AUTOMATIC REGULATION (in Russian). A. A. Feldbaum. *Avtomatika i Telemekhanika*, Moscow, USSR, vol. 10, no. 4, 1949, pp. 249–66.

Discussion

Robert W. Bass, John M. Kopper, and Henry S. McDonald (The Johns Hopkins University, Baltimore, Md.): To anyone who has tried to design relay, or on-off, servomechanisms, the difficulties in both analysis and synthesis are well known. That so much work has been done on linear, continuous-type servos is due to the fortunate circumstance that a neat and complete analytical theory can be worked out easily for this type, largely because little more is involved than a complete examination of the solution to a linear second-order differential equation with constant coefficients. With relay servos the situation is different, as attested to by the many and varied approaches that have been used to date in an effort to analyze the problem so as to make possible a synthesizing technique. When one considers that the majority of servo systems are called upon to respond to steplike changes rather than to sinusoidal changes, it would appear that the approach used by MacColl, Weiss, and Feldbaum (see references of the paper), and now extended by Mr. Hart, is the one most likely to produce good design methods. Using this realistic approach, Mr. Hart has obtained general design curves which should prove of great value in the construction of relay servos for the case of deadbeat operation.

In cases where deadbeat operation is not possible, other criteria for stability are necessary. At Johns Hopkins, in a program being sponsored by Frankford Arsenal, we have been working with other than the deadbeat criterion and have extended Feldbaum's work to include time delay along with hysteresis and velocity feedback. Stability curves similar to those of Mr.

Hart and Mr. Feldbaum have been determined for the case of a resistance-capacitance phase-lead network in the loop. In addition, we have considered the use of relay output voltage feedback, suggested to us by Oscar Wilsker of Frankford Arsenal, who had used it successfully in several experimental relay servomechanisms. A most important result of our work is the analytical determination of a proper combination of position, velocity, and relay output voltage feedback to achieve stability by removing completely the effects of hysteresis and time delay from the system in the sense of step and ramp inputs. From this we have developed a general logarithm for the case of higher-order relay systems. This allows the determination of a set of feedback coefficients (including feedback around the relay) which will exactly stabilize such a system for step and ramp inputs.

We owe Mr. Hart a considerable debt

for making a decidedly forward step in the difficult procedure of going from the analysis of systems to the synthesis of systems.

Rufus Oldenburger (Woodward Governor Company, Rockford, Ill.): The treatment by the author of the effects of various parameters on the performance of relay servo-mechanisms and the study of deadbeat response is most timely. In any thorough study of such relays the parameters discussed by Mr. Hart must be taken into account.

A number of papers[1-7] have been concerned with the intentional use of saturation, such as making a servo travel at full speed (and power) as much of the time as possible. The performance is designed to be deadbeat. For optimum controller-process transients the servo output of the controller must generally travel at full speed at all times, except when the speed should be zero. For the saturation approach the values of controlling functions determine when the servo should travel at full speed in one direction, full speed in the other or stand still.

The theory of relay servos was given considerable impetus by Dr. I. Flügge-Lotz's book.[8]

REFERENCES

1. NONLINEAR TECHNIQUES FOR IMPROVING SERVO PERFORMANCE, Donald McDonald. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 6, 1950, pp. 400–21.

2. A PHASE-PLANE APPROACH TO THE COMPENSATION OF SATURATING SERVOMECHANISMS, Arthur M. Hopkin. *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 631–39.

3. MULTIPLE MODE OPERATIONS OF SERVOMECHANISMS, D. McDonald. *Review of Scientific Instruments*, New York, N. Y., vol. 23, no. 1, 1952, pp. 22–30.

4. THE APPLICATION OF NONLINEAR TECHNIQUES TO SERVOMECHANISMS, Kenneth C. Matthews, Robert C. Boe. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 8, 1952, pp. 10–21.

5. SOME DESIGN CONSIDERATIONS OF A SATURATING SERVOMECHANISM, Perry E. Kendall, James F. Marquardt. *Ibid.*, vol. 9, 1953, pp. 178–87.

6. AN INVESTIGATION OF THE SWITCHING CRITERIA FOR HIGHER ORDER CONTACT SERVOMECHANISMS, Irving Bogner, Louis F. Kazda. *AIEE Transactions*, vol. 73, pt. II, July 1954, pp. 118–27.

7. PREDICTOR SERVOMECHANISMS, L. M. Silva. *Transactions of the IRE, CT-1*, New York, N. Y., March 1954.

8. DISCONTINUOUS AUTOMATIC CONTROL (book), Dr. I. Flügge-Lotz. Princeton University Press, Princeton, N. J.

Time Variation of Industrial System Short-Circuit Currents and Induction Motor Contributions

W. C. HUENING, JR.
ASSOCIATE MEMBER AIEE

CALCULATIONS of the approximate variation with time of industrial power system short-circuit currents were made for a discussion of the circuit-breaker application procedure proposed in an AIEE paper.[1] The results of the calculations indicate that induction motor loads produce contributions to short-circuit currents that do not decay rapidly in every case and perhaps should not be ignored even when present simplified application calculating procedures permit it.[2] In this paper, the method used to determine the approximate short-circuit current time variation is described, and the behavior of induction motors which produce short-circuit current contributions that do not decay rapidly is discussed.

Factors Affecting Short-Circuit Currents

The principal sources of system short-circuit currents are the connected rotating machines, including both motors and generators. When an industrial power system is short-circuited, if the plant has a utility tie, a major contribution to the total short-circuit current is delivered via this tie from the remote-connected utility generators. If synchronous generators are part of the plant power system, they are major contributors. Synchronous and induction motors connected to the plant system, although not considered electric power sources when operating normally, also contribute appreciably to the total short-circuit current under system fault conditions.

The most important factor affecting not only the magnitude but also the time variation of the contribution from each of the rotating machine sources is the impedance between the short circuit and the contributing source. Contributing sources at locations electrically close to the short-circuit point will have the amount of their current contributions limited but little by impedance. Plant synchronous machines are usually electrically close to the short-circuit point, and in calculating their behavior as contributors, accepted procedures for determining terminal short-circuit currents[3] (in which a synchronous machine is represented by a changing impedance that is initially a subtransient value at the start of the short circuit and increases through transient to synchronous, or steady-state, values as time progresses) are followed with modifications to account for series impedances. Induction motors are also often electrically close, and their behavior

as contributors is approximated by simplifying, and modifying to account for added impedance, the accepted equivalent circuit used for calculating their performance.[4]

Contributing sources that are electrically remote will have a high circuit impedance in series to the short-circuit point. Utility generating equipment is considered to be very remote electrically in making calculations of industrial system short-circuit currents. It is assumed that the impedance external to the utility generator terminals is so large with respect to the internal impedance of the generators that the total impedance of the utility circuit to the point of connection in the industrial system is essentially a constant.

Any short-circuit current, and any contribution from a rotating machine source to the short-circuit current, is likely to be asymmetrical and can be analyzed as though it consists of two parts: an alternating component and a direct component. Although the alternating component may in itself consist of a fundamental and several smaller harmonic frequency portions, only the fundamental frequency portion at or near rated frequency for the power system is considered here. Fig. 1 shows an asymmetrical current wave, its components, and rms values for the whole wave and for its alternating component.

Neither the rms value of the alternating component nor the direct component of the contribution from any source remains
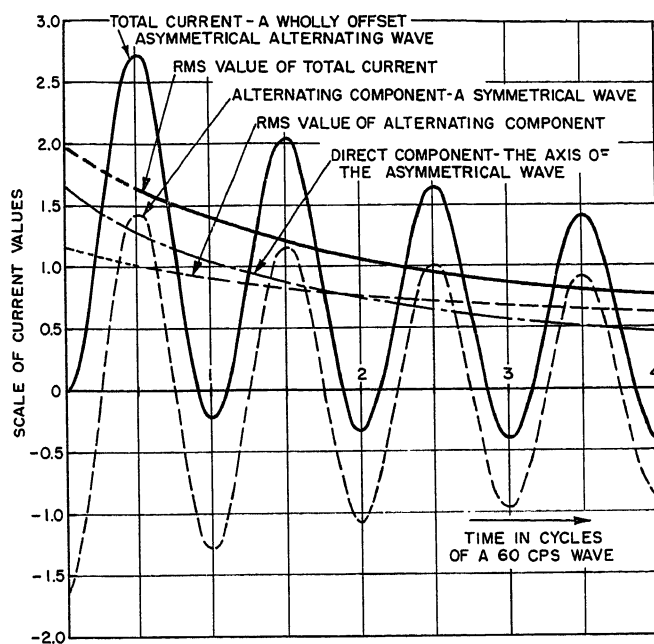
Fig. 1 (above). Structure of an asymmetrical current wave

Fig. 2 (right). Representation of induction motors in short-circuited systems

---
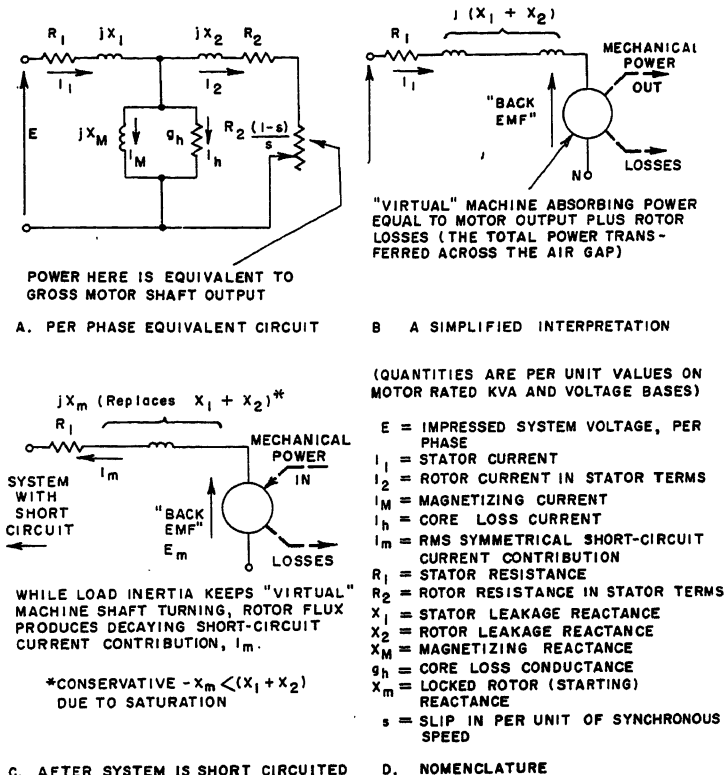
at a constant magnitude as time passes from the initial instant, when the short-circuit current starts to flow. Usually, the magnitude of each component decays nearly exponentially, and a time constant can be found to represent each decay pattern. Since time constants are determined by circuit reactance and resistance values, external impedance in the path from a contributing machine to the short circuit affects not only the magnitude but also the decay pattern of both the alternating and direct components of the short-circuit current contribution.

Basis of Calculating Procedure for Time Variation

The basic approach to determining the time variation of each current component by this method is to evaluate the initial magnitude of the component and its initial exponentantial decay and to obtain an approximate value for the limiting final magnitude. The decay of each current component from the initial current magnitude toward the approximate final magnitude proceeding exponentially according to its initial pattern is assumed to be nearly correct for several cycles after the start of a short circuit.

The calculation of the approximate time variation of a total system short-circuit current following the basic approach consists of the following steps:

1. Development of the equivalent circuit.
2. Solution of the circuit for initial or

zero time ($t=0$), contributions from each contributing source.

3. Representation of each separate source in an "individual" circuit, in which the initial value of source current flows alone through a single external impedance. This definition of an individual circuit is used throughout the paper.

4. Determination of the time variation of the decaying alternating and direct components of each contribution separately in its individual circuit.

5. Summation of the individual contributions to obtain the variation of the symmetrical alternating and the direct components of the total short-circuit current, and computation from these of an asymmetrical rms total short-circuit current time variation.[2]

EQUIVALENT CIRCUIT

Following usual short-circuit calculating procedures, a solution is sought for the worst fault condition, that of a 3-phase "bolted" short circuit. The circuit symmetry during a bolted 3-phase fault permits a "per-phase" calculation.

An equivalent circuit is drawn, using the system 1-line diagram as a guide. Each static circuit element is represented by a single series impedance and each rotating machine by a single impedance in series with a voltage source. Again following usual calculating procedures, all shunt branches (magnetizing reactances, line capacitive reactances, etc.) are omitted. Values are inserted for each impedance after conversion to per-unit quantities on a common base kilovolt-

amperage and voltage. Because both affect time constants, resistance and reactance values are given for each impedance. The impedance substituted for each contributing source is selected so that a "conservative" value will be obtained for its initial symmetrical contribution. A conservative value for a short-circuit current is one calculated so that any error gives a current value larger than actual. A circuit breaker applied on the basis of the calculation should never be in trouble for it should never be required to carry a current quite as high as the calculated one. Equivalent circuit impedances thus tend toward minimum values.

The utility circuit, if one exists, is represented by a reactance, corresponding to the maximum magnitude of short-circuit kilovolt-amperes that can be delivered from the utility system to the point of connection to the industrial system in series with a voltage source. This maximum available fault kilovolt-amperes can generally be obtained from utility engineers. A resistance may be assigned to complete the equivalent impedance if a value for the reactance-resistance ratio is also available from utility engineers, or can be assumed.

Groups of similar rotating machines are often lumped together and treated as one. Synchronous machinery is represented by stator resistance in series with subtransient reactance and a voltage source. Subtransient reactance is selected because it is the reactance quantity, of

those associated with synchronous machine behavior, that corresponds to initial current after a sudden change like the application of a fault. Use of this smallest of the synchronous machine reactances in series with a voltage source in the system equivalent circuit gives a conservative value for the initial contribution to a system short circuit.

The single impedance representing each induction motor, or group of induction motors, in the system equivalent circuit is a simplification of the complete equivalent circuit for an induction motor shown in Fig. 2(A). The first step in the simplification is the omission of the magnetizing and core loss branches yielding a simplified equivalent circuit shown in Fig. 2(B). The power delivered to the right-hand circuit element in Fig. 2(A) equals the motor shaft power output plus losses associated with the shaft rotation. The shaft output is mechanical energy which keeps the motor's rotating load in motion. In Fig. 2(B) the right-hand elements of Fig. 2(A) have been replaced by a "virtual" rotating machine absorbing a total power equal to the power transferred across the air gap of the motor and, of course, there must be a corresponding voltage across the virtual machine terminals. The virtual machine voltage may be regarded as a "back-electromotive force" which represents the voltage gen-

erated by the rotor flux cutting the stator conductors.

If a short circuit appears on the system connected to the motor terminals, as in Fig. 2(C), the system may no longer be able to deliver electric power to the motor. Yet the rotating mechanical load does not stop immediately; its inertia kept it turning and keeps the induction motor shaft turning also. The rotor speed does not change appreciably for a short period after the system short circuit occurs because the system loses comparatively little energy. (Note that short-circuit currents are low power-factor currents.)

The virtual machine back-electromotive force does not disappear either. Since it cannot change instantaneously, the flux linking the induction motor rotor conductors takes time to diminish from its initial magnitude during the period after the occurrence of the short circuit. This trapped rotor flux turns with the shaft and produces a 3-phase voltage at the motor terminals which is represented by the back-electromotive force of the equivalent circuit. With this voltage present, a current contribution flows from the induction motor to the system short circuit.

The short-circuit current contribution may be an offset alternating wave which has a direct, as well as an alternating, component. The direct component may

be regarded as though it were being produced by the decay of trapped flux which linked only the stator windings at the instant of the system short circuit. This explains the relationship shown by the equations in Appendix I that the induction motor short-circuit current alternating-component decay is governed by rotor resistance, while the direct-component decay is governed by stator resistance.

The reactance used in the system equivalent circuit for the induction motor under short-circuit conditions is equal to the locked-rotor reactance determined during motor locked-rotor tests. In other words, the induction motor with a terminal short circuit is assumed to produce an initial current equal to the full voltage starting current.

System short-circuit locations to be investigated are selected and marked on the 1-line diagram and the equivalent circuit. Fig. 3(A) shows a 1-line diagram of a simple hypothetical system with a short-circuit point located, and Fig. 3(B) shows an equivalent circuit for the system.

SOLUTION FOR INITIAL CONTRIBUTIONS

Once the equivalent circuit has been completely drawn, with each source of short-circuit current represented by a series impedance and a voltage source, and each line or transformer represented
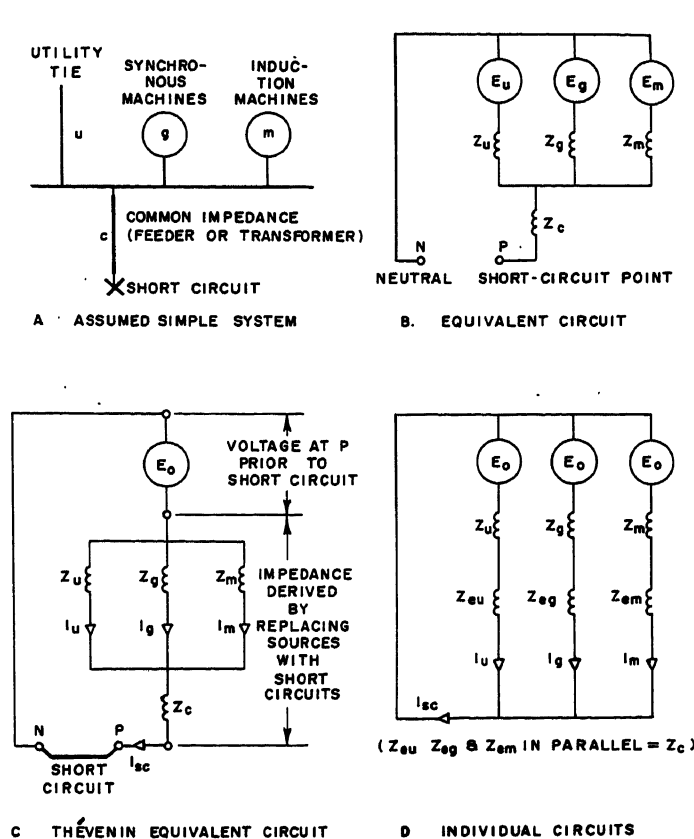
Fig. 3. Method applied to simple system
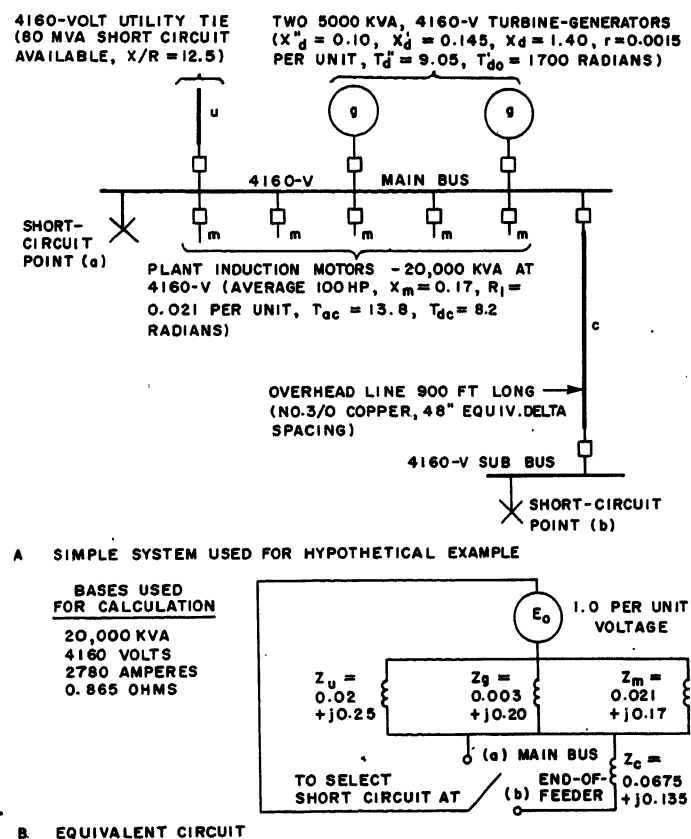
Fig. 4. Example system and equivalent circuit

Fig. 5. Example initial currents and individual circuits
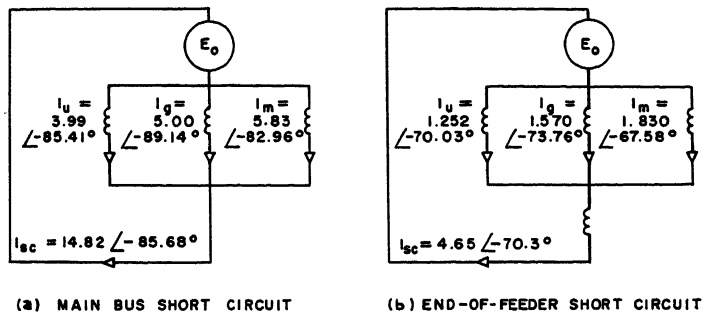
A   INITIAL RMS ALTERNATING COMPONENTS OF SOURCE CURRENTS

(a) MAIN BUS SHORT CIRCUIT

$I_u = 3.99 \angle -85.41°$ $I_g = 5.00 \angle -89.14°$ $I_m = 5.83 \angle -82.96°$

$I_{sc} = 14.82 \angle -85.68°$

(b) END-OF-FEEDER SHORT CIRCUIT

$I_u = 1.252 \angle -70.03°$ $I_g = 1.570 \angle -73.76°$ $I_m = 1.830 \angle -67.58°$

$I_{sc} = 4.65 \angle -70.3°$
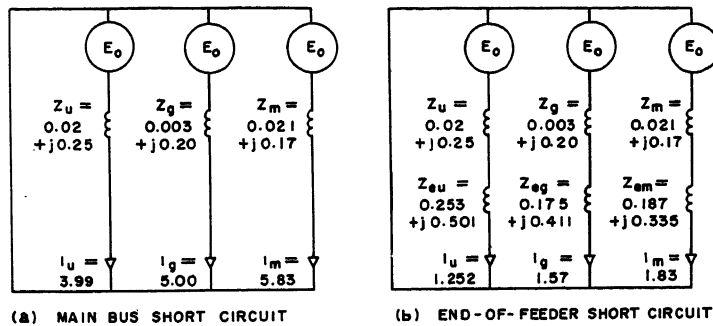
B.   INDIVIDUAL SOURCE CIRCUITS TO SOLVE FOR TIME VARIATIONS

(a) MAIN BUS SHORT CIRCUIT

$Z_u = 0.02 + j0.25$ $Z_g = 0.003 + j0.20$ $Z_m = 0.021 + j0.17$

$I_u = 3.99$ $I_g = 5.00$ $I_m = 5.83$

(b) END-OF-FEEDER SHORT CIRCUIT

$Z_u = 0.02 + j0.25$ $Z_g = 0.003 + j0.20$ $Z_m = 0.021 + j0.17$

$Z_{eu} = 0.253 + j0.501$ $Z_{eg} = 0.175 + j0.411$ $Z_{em} = 0.187 + j0.335$

$I_u = 1.252$ $I_g = 1.57$ $I_m = 1.83$

Fig. 6. Short-circuit current at main bus of example

by a series impedance, then the Thévenin (or Helmholtz) theorem is used to facilitate the solution for the short-circuit current contributions to the chosen short-circuit points. One reference[5] states Thévenin's theorem as follows:

"In the alternating-current steady state and for any given single frequency, any network containing only linear passive elements and constant vector-voltage or current sources is, when viewed from any given pair of terminals, indistinguishable from a simple vector-voltage source consisting of a constant vector electromotive force in series with a constant impedance. The vector electromotive force is equal to the open-circuit voltage across the given pair of terminals, and the series impedance is equal to the impedance is the given network as viewed from the same terminals with all sources (meaning "source voltages") replaced by connections of zero impedance."

The solution of an equivalent circuit for initial currents may be carried out as a steady-state calculation when the impedances are chosen to match initial conditions, as previously described. Transient decays are considered later. The given pair of terminals will be. the point at which the short-circuit current is to be investigated, and the system neutral point. The open-circuit voltage is the voltage appearing between these two points before the short circuit is initiated

and it is usually about equal to the system nominal voltage or 1.0 per unit. The impedance in series with this voltage in the Thévenin simplification is comprised of all the impedances of the system equivalent circuit, the voltage sources having been replaced with zero impedance connections. Fig. 3(C) shows the Thévenin equivalent for the simple illustrative system.

The equivalent network is solved for current contributions through all source elements, using either an a-c network calculator or a slide rule. The source contributions must be obtained in complex form.

The Thévenin theorem analysis gives the correct solution of the system equivalent circuit for the initial value of the total short-circuit current, but the current contributions found flowing in the various source circuits of the Thévenin equivalent circuit do not represent the total currents that may flow in the corresponding machine windings under actual system short-circuit conditions. Thévenin circuit current through source impedances represent only the contributions that the source machines deliver to the total short-circuit current. Other components of current may also be flowing in the machine windings associated, e.g. with power exchanges or static loads. Because these other components are usually relatively small and they are not in phase with the

contributions from the machines to the total short-circuit current, for the purpose of calculating approximate time variations each machine contribution determined from the Thévenin circuit is treated as though it alone is flowing through the machine windings.

AN INDIVIDUAL CIRCUIT FOR EACH SOURCE

The complete Thévenin equivalent circuit is now disassociated into a number of separate individual circuits with each carrying only the initial current contribution of one of the original contributing sources. All these individual circuits are treated as though they delivered their contributions separately to the short-circuit point where the contributions add up to equal the total short-circuit current. For the simple hypothetical system, this synthesis is shown in Fig. 3(D).

The whole impedance of each one of these individual circuits is equal to the Thévenin circuit voltage divided by the initial current contribution that the circuit carries and, expressed as a complex number, has both resistance and reactance terms. The impedance of the contributing source machine is then subtracted from the whole impedance of the individual circuit to obtain an effective external resistance and reactance that acts in series with the machine impedance.

Time Variation of Contributions From Each Individual Circuit

The time variation of the current in each individual circuit is now calculated as though that circuit were isolated, using appropriate equations from those listed in Appendix I. The equations in Appendix I are modified from equations for computing total machine currents during terminal short circuits. The modifications are made only to take into account the external impedance added in the individual circuits. The equations for d-c components of each contribution are written to give the maximum possible values, based on the assumption that each contribution is a fully offset alternating wave at its initiation.

Neither the alternating nor direct component of a source current will be maintained at its initial value. Usually, both will decay from their initial values with decay patterns determined by the machine time constants modified by the effects of the external resistance and reactance. The decays proceed toward final values based on what would be obtained if the machines actually were acting alone in their individual circuits.

Values for both alternating and direct components of each contribution are obtained at several instants in time from 0 to 4 or 5 cycles. Note that both components of all contributions should be evaluated for the same chosen instants so that summations can be made for those instants.

Determining the Total Short-Circuit Current Values

At each chosen instant, the alternating and direct components of the total short-circuit current are found by adding up all the alternating and direct components of every contribution as they were computed for that instant. An rms value at each instant can then be computed from the total alternating and direct components by standard procedures.[2] (This rms current evaluated at an instant is actually a very close approximation for the true rms value over 1 cycle of the current wave for which the instant is a mid-point. It is equal to the square root of the sum of the squares of the rms value of the symmetrical alternating component and the magnitude of the direct component obtained at the same instant.)

Example of the Calculation

The method of calculation described is used to determine short-circuit current-time variations of a hypothetical simple circuit as an example. The circuit includes utility tie, synchronous generators and induction motors connected to a 4,160-volt main bus with the arrangement and data given in Fig. 4(A). Two short-circuit points are investigated: one at the main bus, a, and one at the far end of a rather long feeder originating at the bus, b. The equivalent circuit of the

Table I. Equations for Approximate Variations with Time of Components of Current Contributions to Short Circuits of Example

Current Component	Short Circuit at a	Short Circuit at b
Utility alternating rms symmetrical, $i_u =$	3.99	1.252
Utility direct, $i_{ud} =$	$5.64\,\epsilon^{-t/12.5}$	$1.77\,\epsilon^{-t/2.76}$
Synchronous generator alternating rms symmetrical, $i_g =$	$1.55\,\epsilon^{-t/9.05}$	$0.179\,\epsilon^{-t/10.44}$
	$+3.09\,\epsilon^{-t/176}$	$+1.077\,\epsilon^{-t/371}$
	$+0.357$	$+0.311$
Synchronous generator direct, $i_{gd} =$	$7.07\,\epsilon^{-t/65.7}$	$2.22\,\epsilon^{-t/2.44}$
Induction motor alternating rms symmetrical, $i_m =$	$5.83\,\epsilon^{-t/13.2}$	$1.830\,\epsilon^{-t/41}$
Induction motor direct, $i_{md} =$	$8.24\,\epsilon^{-t/8.2}$	$2.585\,\epsilon^{-t/2.42}$

t is in radians, and $t = 2\pi t'$ if t' is in cycles.

Fig. 7 (below). Short-circuit current at end-of-feeder of example

Fig. 8 (right). One-line diagram of an extensive large industrial power system

NOTE* ON BASES USED FOR EXAMPLE—
"SHORT CIRCUIT CURRENT" = 4.72 PER UNIT
—SEE REFERENCE 1 OF THE PAPER).

RMS TOTAL ASYMMETRICAL SHORT CIRCUIT CURRENT

RMS SYMMETRICAL SHORT CIRCUIT CURRENT (ALTERNATING COMPONENT OF TOTAL CURRENT)

SYMMETRICAL INDUCTION MOTOR CONTRIBUTION

SYMMETRICAL SYNCHRONOUS GENERATOR CONTRIBUTION

SYMMETRICAL UTILITY CONTRIBUTION

PER UNIT CURRENT WITH "SHORT CIRCUIT CURRENT" = 1.0*

TIME IN CYCLES (OF A 60 CPS WAVE) AFTER START OF SHORT CIRCUIT

GEN = TURBINE-GENERATOR
SM = SYNCHRONOUS MOTORS
IM = INDUCTION MOTORS

UTILITY TIE 34.4 KV

Fig. 9. Short-circuit current at one of several medium-voltage busses in an extensive large industrial power system

Fig. 10. Short-circuit current at main bus of example compared with simplified calculation results

example, with all impedances expressed in per unit on a 20,000-kva 4160-volt base, is shown in Fig. 4(B).

The circuit is solved for initial current contributions through the source branches with short circuits at a and b yielding results noted in Fig. 5(A) at a and b. Individual circuits are synthesized for each contributing source. Results are shown in Fig. 5(B) at a and b. In a each source suffers a terminal short circuit, so the individual circuits contain source internal impedances only. In b total impedances for the individual circuits are determined from initial contribution currents, and each is separated into machine plus effective external impedance.

Equations for the time variations of the alternating and direct components of each contribution are taken from Appendix I with proper numerical values inserted, and are shown in Table I. For a short circuit at a, the sources suffer terminal short circuits and the external resistance and reactance terms in the equations are equal to zero. For the short circuit at b, the equations have included the influence of the external effective impedance in changing both magnitude and decay time constants.

The total short-circuit current alternating and direct components are obtained for several instants by summing all the calculated alternating and direct com-

ponents of the contributions at the several instants. The rms value at each instant is then determined.

The results of the calculations for the total current-time variations for short circuits at a and b are shown in Figs. 6 and 7 respectively. In these figures, the currents have been normalized by dividing by the per-unit short-circuit current calculated at a and b respectively, using the recently proposed application procedure.[1] (Short-circuit current for $a = 13.0$ per unit, and for $b = 4.72$ per unit.) The composition of the rms symmetrical alternating components of the total short-circuit current is shown by its division into portions associated with each of the three sources, so that the decay of the separate components may also be observed.

Considerably more work is required to obtain the total short-circuit current-time variation of more complicated systems. For a second illustration, consider the system of Fig. 8, with a short-circuit location as shown at one of several medium voltage busses in an extensive industrial power system. Over 25 different individual source circuits were solved before a summation could be made. In Fig. 9 the results of an investigation of this circuit are plotted, and the composition of the rms symmetrical component of the total current is again displayed by

division into separate contributions from the different kinds of sources.

Usefulness of the Approximate Calculation

It is evident that a short-circuit current calculated according to the methods described will be an approximation only. If, however, the relationship between the approximate current and the actual current that might flow at the desired point in the physical system is known, the approximate value then becomes a useful tool. Since the original purpose of the calculations was to investigate circuit-breaker applications, attempts were made to insure that the calculated results would be higher than actual. If a circuit breaker is adequate for application when the short-circuit current equals a calculated value known to be larger than actual, the circuit breaker should certainly be adequate for the actual short-circuit current.

Basic assumptions definitely tending to make the calculations yield conservative, above-actual results are repeated here.

1. Minimum series impedances are used to represent source machinery in the equivalent circuit.

2. D-c components are based on the complete offset of every initial alternating component, a condition actually improb-
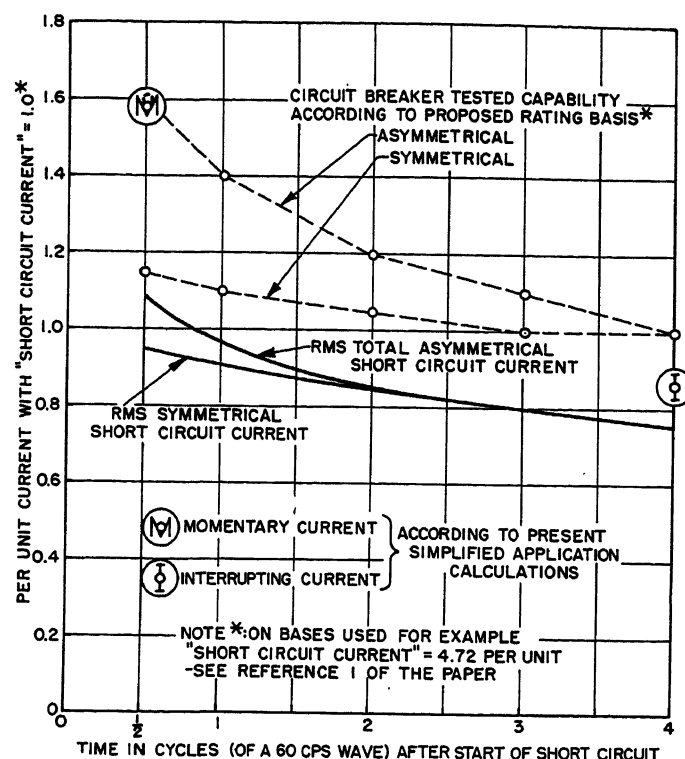
Fig. 11. Short-circuit current at end-of-feeder of example compared with simplified calculation results
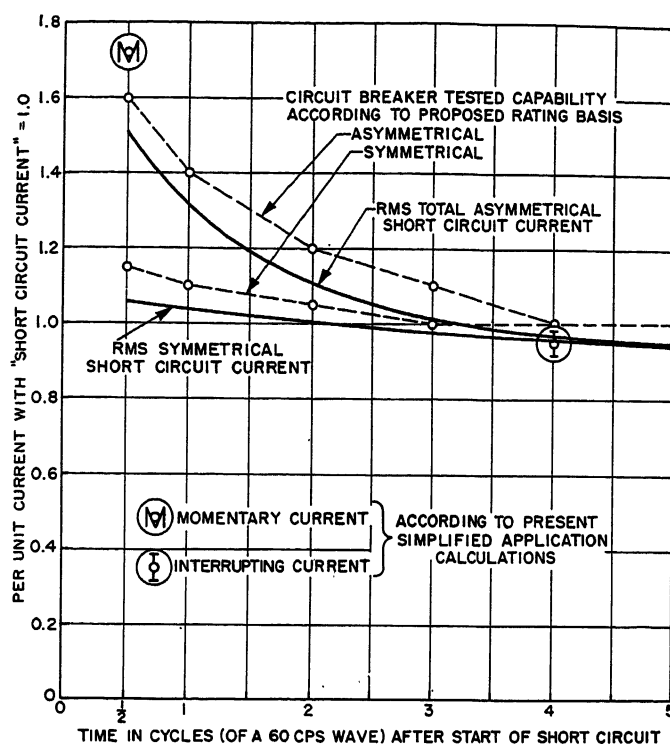
Fig. 12. Short-circuit current at one of several medium-voltage busses in an extensive large industrial power system calculation results

able for it depends on all initial current contributions being in phase.

Other inaccuracies of the method do not so definitely tend to produce conservative results, yet apparently they do not destroy the conservatism. One such inaccuracy is introduced when the action of several machine contributions flowing through a common impedance is treated as though each of the several contributions acted separately, starting at an initial value to follow a simple exponential decay. This is correct only if the exponential decays of the several contributions all have equal time constants. When all contributions do not have the same time constants, there are interactions that cause changed decay patterns among the several contributions.

To illustrate the inaccuracy, refer to part *b* of the example, an application of the approximate method to a simple case when different contributions flow through a common impedance with approximate results plotted in Fig. 7. In the calculation, the effect of the common impedance is approximated by inserting in series with each of the sources an effective external impedance, chosen so that the parallel combination of the three external impedances equals the original single common impedance, and so that the initial currents are correct. The decay patterns of the alternating components of the induction motor and synchronous generator short-circuit current contributions are

modified to account for the effective external impedances. The effective external impedance in the utility circuit does not make its alternating component vary with time in the approximate method, though a time variation would probably occur if the system were real.

The approximate method illustrated ignores the fact that the voltages at the junction points between internal and effective external impedances in the individual source circuits must all be equal at all times since they represent actual connections to a single common point, the main system bus. This condition, met only at zero time, is no longer met after a small increment of time passes, and the error introduced increases as time progresses. To reduce the inaccuracy, the effective external impedances of the individual circuits should also vary with time.

As soon as some small increment of time has passed, the induction motor alternating contribution drops from its initial value with an incremental change much greater than that of the alternating contribution from the utility which at first actually changes slowly, if at all. To keep the same bus voltage in the three individual circuits, the effective external impedance in the induction motor circuit must increase so that the smaller current will produce a voltage equal to that at the junction points in the other circuits. If the effective external impedance in the

induction motor circuit increases, then the effective external impedances in the other circuits must also change in order to keep the parallel combination at a constant value equal to the common static impedance, and without question the effective external impedance in the utility circuit will decrease. The decrease in utility circuit impedance indicates that the utility contribution, instead of remaining a constant, should actually be increasing. Since the synchronous machine circuit has an intermediate time constant value, its effective external impedance may remain constant or change either way, but probably with a smaller increment than the change in either of the other circuit impedances.

The increase in effective external impedance of the induction motor individual circuit has two effects: the time constant of the alternating component tends to increase and the magnitude tends to decrease. To some extent these two effects offset each other, and the induction motor contribution considering interaction may be much the same as the approximate value calculated ignoring interaction.

As time passes, magnitude and time constant changes, which should be but are not considered, cause larger and larger errors in the values computed by the approximate method described. Sample solutions taking into account the interaction were checked against corresponding

approximate solutions, and the checks indicate that the approximate method gives results accurate to about 5 per cent at 4 cycles after the short circuit starts. The method is thus considered reasonably accurate to 4 or 5 cycles.

The accuracy of the approximate calculated decay of the d-c contributions from the sources is probably poorer than that of the alternating decay. The exact behavior of a d-c component in an a-c circuit is difficult to compute for a check. The values obtained for d-c contributions by the approximate method will, however, probably give a d-c component of the whole short-circuit current that is conservative, because of the original assumption that every source current is initially offset to the maximum possible extent.

The approximate method described may supplement, though it is not intended to replace, simplified application calculations. It has some drawbacks, including the following: considerable time is required to obtain an answer; results cover reasonably well only a period of 4 or 5 cycles after the start of a short circuit; and very complete information about power system machinery impedances and time constants must be gathered. Present and proposed calculating procedures do not require information about resistances, synchronous machine synchronous reactances, or any time constants as this method does. It may be necessary to obtain this required information directly from the machinery manufacturer, although some approximate information has been given in Appendix II.

Comparison With Standard Circuit-Breaker Application Calculations

For general interest, the simplified application procedures now standard[2] and recently proposed[1] are applied to the circuits used as examples. Figs. 10, 11, and 12 show replots of the results of the illustrative examples which have superimposed the momentary and interrupting duties calculated according to present procedures, and the breaker test characteristic required according to proposed procedures. The momentary duty, analogous to the rms value of the current over the first cycle, is shown at 1/2 cycle after the start of the short circuit. The interrupting duty, calculated for an 8-cycle breaker, is shown at 4 cycles based on the assumption that 8-cycle circuit breakers may start to part contacts at that time.

Induction Motor Behavior

When the electric supply system is short-circuited close to the terminals of an induction motor, of course the current contributed by the induction motor to the system short circuit is equal to the motor terminal short-circuit current. When the system short-circuit point is some distance away, the current contributed by the induction motor differs from terminal short-circuit current, for it is affected both by the impedance of the circuit from its terminals to the short-circuit point and by outside influences associated with dissimilar machines whose contributions follow the same path to the short-circuit point.

High-speed induction motors and ones with large horsepower ratings have terminal short-circuit currents that take relatively long times to decay. This is shown by the rising curves in the graph of induction motor time constant trends, Fig. 13. If it is assumed that a 10-percent error in the total current at contact-parting times of 4 cycles, because of ignored induction motor current contributions, is intolerable, and if it is calculated that the initial induction motor contribution is 30 per cent of the total, the tolerable error is exceeded if about one-third of the induction motor contribution remains after 4 cycles. This corresponds to a time constant for the induction motor alternating component decay of about 4 cycles or more. (In one time constant, the decaying quantity drops to about 37 per cent of its initial value.)

Motors having a time constant of 4 cycles or more usually have rated voltages of 2,300 volts or higher. At corresponding system voltages, cable impedances are small enough to be considered insignificant, and motor contributions have the same decay patterns that terminal short-circuit currents have. To summarize, at system voltages of 2,400 volts or higher if a short circuit occurs in a location where there are induction motors having large horsepower ratings connected at system voltage, and the motor contribution to the initial (momentary) short-circuit current is appreciable, it is very likely that the still-present motor contribution to the duty during a circuit-breaker interruption cannot be considered negligible. This may be important for 8-cycle circuit breakers tripped instantaneously and, of course, is even more important if the breaker is faster.

External impedance associated with transformers or lengths of feeder, effectively in series between induction motors and a system short circuit, affects both magnitude of initial contribution and its rate of decay. The external impedance reduces the initial contribution from each motor to a value smaller than the terminal short-circuit current. At the same time, the external impedance increases the time constant of the alternating component decay so that the reduced contribution, although smaller initially, lasts longer. Depending on the nature of the external impedance, the time constant of the direct component decay may or may not be changed. For example, where the reactance-resistance ratio of the external impedance is lower than that of the motor, the reactance-resistance ratio of the whole circuit is reduced and
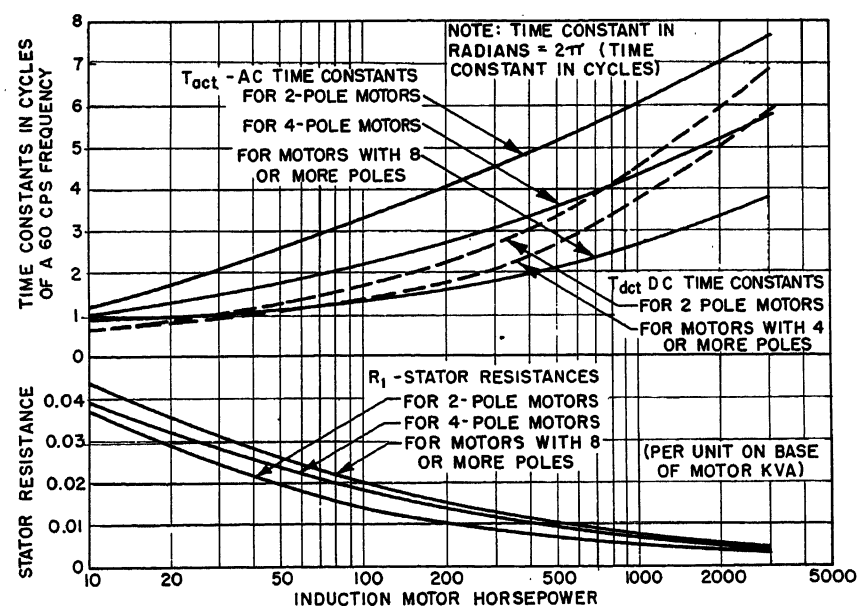
Fig. 13. Average trends of terminal short-circuit time constants and stator winding resistance values for 60-cycle-per-second general-purpose induction motors

X_d'' Range	X_d' Range	X_d Range	r Range	T_{dt}'' Range	T_{do}' Range
		Turbine Generators, 20,000 Kva and Below			
0.08–0.13	0.10–0.20	1.00–1.50	0.0015–0.003	5–15	375–1,900
		6–14 Pole Synchronous Motors over 250 Horsepower, 2,300 Volts and Above			
0.10–0.25	0.15–0.40	1.00–1.75	0.005–0.015	10–20	375–1,200
		Synchronous Motor Less Than 250 Horsepower, Low Voltage			
Assume 0.15	Assume 0.25	Assume 1.20	Assume 0.02	Assume 15	Assume 500

so is the direct component time constant. To summarize, at any voltage level if induction motors are separated by impedance from the point of fault, yet the initial (momentary) contribution from the motors to a system short circuit is appreciable because there are many motors, it is very likely that the still-present contribution to the duty during a circuit-breaker interruption cannot be considered negligible.

If the path associated with the external impedance is shared with synchronous machine and utility contributions, an interaction may occur which modifies the effects of simply having impedance inserted in series with the induction motor. Possibly the interaction may cause the alternating component of the induction motor short-circuit current contribution to decay with a different pattern.

The effect of external impedance in prolonging induction motor contributions to system short-circuit currents is illustrated by the example used in describing the short-circuit calculations. In Fig. 6 conditions are shown during a bus short circuit, which produces induction motor currents like those during a terminal short circuit. The alternating component of the induction motor contribution at 4 cycles is reduced to 20.2 per cent of its 1/2-cycle value, and is only 7.15 per cent of the total symmetrical short-circuit current. In Fig. 7 conditions are shown during a short circuit at the end of a long feeder far away from the bus. The alternating component in the induction motor contribution at 4 cycles is 58.5 per cent of its 1/2-cycle value and is still 20 per cent of the total symmetrical short-circuit current. This is a hypothetical case, but it can be seen that the result of a study on an actual system, as shown in Fig. 9, follows the same pattern.

Summary

Short-Circuit Calculating Procedure

The calculating method described and illustrated in this paper should give results that will prove satisfactory in many cases where the approximate time variation during a few cycles of a short-circuit current is desired. The method is not suggested as a replacement for simplified application calculations for power circuit breakers, but it may supplement the simpler calculations in some circuit-breaker application problems. Though time-consuming, the calculations can be carried out using a log-log duplex slide rule.

Behavior of Induction Motor Short-Circuit Current Contributions

In cases where induction motor contributions are a significant portion of a short-circuit current during the first cycle, they may also be significant during a later circuit-breaker interruption. Formerly assumed to decay so rapidly that they soon disappeared, the induction motor contributions actually may diminish more slowly and remain appreciable for several cycles, particularly when there is some system impedance between the motor terminals and the short circuit.

Appendix I. Calculation Details

Equations

For utility contribution components

Rms alternating: $i_u = I_u = \left| \dfrac{E_0}{Z_u + Z_{eu}} \right|$

Direct: $i_{ud} = \sqrt{2} I_u \epsilon^{-t/T_u}$

For synchronous machine (motor and generator) contribution components

Rms alternating: $i_g = E_0 \times$

$$\left(\left| \frac{1}{Z_d'' + Z_{eg}} - \frac{1}{Z_d' + Z_{eg}} \right| \epsilon^{-t/T_d''} + \right.$$

$$\left| \frac{1}{Z_d' + Z_{eg}} - \frac{1}{Z_d + Z_{eg}} \right| \epsilon^{-t/T_d'} +$$

$$\left. \left| \frac{1}{Z_d + Z_{eg}} \right| \right)$$

Direct: $i_{gd} = \sqrt{2} \left| \dfrac{E_0}{Z_{av}'' + Z_{eg}'} \right| \epsilon^{-t/T_a}$

For induction motor contribution components

Rms alternating:

$$i_m = I_m \epsilon^{-t/T_{ac}} = \left| \frac{E_0}{Z_m + Z_{em}} \right| \epsilon^{-t/T_{ac}}$$

Direct: $i_{md} = \sqrt{2} I_m \epsilon^{-t/T_{dc}}$

Nomenclature

$E_0 =$ voltage at short-circuit point just before the start of short-circuit current flow

$E_u =$ voltage behind utility impedance

$E_g =$ voltage behind synchronous machine (motor or generator) subtransient impedance

$E_m =$ induction motor back-electromotive force

$I_u =$ initial rms value of alternating component of utility contribution

$I_g =$ initial rms value of alternating component of synchronous motor contribution

$I_m =$ initial rms value of alternating component of induction motor contribution

$i_u =$ time varying rms value of alternating component of utility contribution

$i_g =$ time varying rms value of alternating component of synchronous machine contribution

$i_m =$ time varying rms value of alternating component of induction motor contribution

$i_{ud} =$ time varying value of direct component of utility contribution

$i_{gd} =$ time varying value of direct component of synchronous machine contribution

$i_{md} =$ time varying value of direct component of induction motor contribution

$Z_u = R_u + jX_u =$ impedance of utility system

$Z_g = Z_d'' = r + jX_d'' =$ subtransient impedance of synchronous machine

$Z_m = R_1 + jX_m =$ locked rotor impedance of induction motor, generally taken equal to the reciprocal of the per-unit starting current

$Z_d' = r + jX_d' =$ transient impedance of synchronous machine

$Z_d = r + jX_d =$ synchronous impedance of synchronous machine

$Z_{av} = r + jX_{av}''$

$Z_{eu} = R_{eu} + jX_{eu} = (E_0/I_u) - Z_u =$ external effective impedance of utility individual circuit

$Z_{eg} = R_{eg} + jX_{eg} = (E_0/I_g) - Z_g =$ external effective impedance of synchronous machine individual

$Z_{em} = R_{em} + jX_{em} = (E_0/I_m) - Z_m =$ external effective impedance of induction motor individual circuit

$X_{av}'' = (2X_d'' X_q'')/(X_d'' + X_q'')$, where $X_q'' =$ quadrature-axis subtransient reactance

$r =$ stator resistance of synchronous machine

$R_1 =$ stator resistance of induction motor

$R_2 =$ rotor resistance of induction motor

$T_u = (X_u + X_{eu})/(R_u + R_{eu}) =$ time constant of decay of direct component of utility contribution, radians

$T_d'' = T_{dt}'' \dfrac{X_d'(X_d'' + X_{eg})}{X_d''(X_d' + X_{eg})} =$ time constant of decay of subtransient alternating component of synchronous machine contribution, where $T_{dt}'' =$ subtransient time constant for synchronous machine terminal short-circuit current

$T_d' = T_{do}' \dfrac{X_d' + X_{eg}}{X_d + X_{eg}} =$ time constant of decay of transient alternating component of synchronous machine contribution, where $T_{do}' =$ open-circuit transient

time constant of synchronous machine

$T_a = (X_{av}'' + X_{eg})/(r + E_{eg}) = $ time constant of decay of direct component of synchronous machine contribution, radians

$T_{ac} = T_{act} + (X_{em}/R_2) = $ time constant of decay of alternating component of induction motor contribution in radians, where $T_{act} = X_m/R_2 = $ time constant of alternating component of induction motor terminal short-circuit current

$T_{dc} = (X_m + X_{em})/(R_1 + R_{em}) = $ time constant of decay of direct component of induction motor contribution, radians

Time constants given in radians are 2π times the corresponding time constants in cycles

Appendix II. Machine Data

For approximate information on certain synchronous machine impedances and time constants see Table II. Values of induction motor locked rotor reactance X_m average about 0.16 per unit. In the absence of more definite information, this value is usually assumed for motors rated 2,300 volts and above. For low-voltage motors, the connection wiring adds appreciable impedance that need not be computed accurately for every case. An average allowance is made for the effects of connection wiring by using an increased value for X_m of 0.25 per unit. For approximate values of induction motor stator resistances and short-circuit time constants see Fig. 13.

References

1. A New Basis for Rating Power Circuit Breakers, AIEE Committee Report. *AIEE Transactions*, vol. 73, pt. III, April 1954, pages 353–67.

2. Methods for Determining the Rms Value of a Sinusoidal Current Wave and a Normal-Frequency Recovery Voltage and for Simplified Calculation of Fault Currents. *Publication No. C37.5—1953*, American Standards Association, New York, N. Y. Dec. 17, 1953.

3. An Introduction to Power System Analysis (book), F. S. Rothe. John Wiley & Sons, Inc., New York, N. Y., 1953, chap. 5.

4. The Nature of Polyphase Induction Machines (book), P. L. Alger. *Ibid.*, 1951, chap. 5.

5. Electric Circuits (book), M.I.T. Electrical Engineering Staff. *Ibid.*, 1940, pp. 469–70.

Discussion

R. L. Webb (Consolidated Edison Company of New York, Inc., New York, N. Y.): This paper will be helpful to a special Working Group which has been set up by the AIEE Committee on Switchgear. This group has been requested to take the recommendation given in reference 1 of the paper and determine whether the proposed short-circuit characteristic fits the needs of different systems with reasonable conservatism.

The Working Group must, of course, consider the effects of induction motors on the total short-circuit current which must be handled by the power circuit breakers, and we have found that the suggestions made by Mr. Huening are helpful in doing this where large industrial plants or generating station auxiliary systems are involved in the calculations.

C. A. Woodrow (General Electric Company, Schenectady, N. Y.): Publication of the material in this paper is both important and timely. Fortunately, the preliminary studies by Mr. Huening were made available in 1953 to the Working Group on Methods of Rating which prepared reference 1 of the paper. In consideration of this induction motor problem which Mr. Huening's paper now so clearly evaluates for typical industrial installations, reference 1 states: "As circuit breaker and relay speeds become faster, the present practice of neglecting the induction motor contribution at time of contact parting is no longer conservative." For standard application, the proposed rating basis (reference 1) attempts to solve this problem by conservatively discounting the induction motor contribution rather than by completely disregarding that contribution at contact parting. That discounting is accomplished by using "effective" induction motor reactances which are appropriately higher than the "motor locked rotor reactance X_m"—specifically, 50 per cent higher for all except big, high-speed motors.

Mr. Huening's Figs. 10 and 12 clearly illustrate conditions which are reasonably covered by the "Guide for the Simplified Calculation of Fault Currents" which is included in the proposed rating basis. This guide recognizes that occasionally conditions will arise where the use of a more

rigorous—but necessarily more complicated—procedure (such as that now made available by Mr. Huening) may be justified "to take advantage of the minor margin that may exist between the rated breaker characteristic and the actual system characteristic." For applications such as those of Fig. 10 or 12, I trust Mr. Huening will agree that this margin is so minor that it would seldom be worth while to tackle the more complicated, time-consuming method he has made available. In this connection, it is interesting to note that for modern power station auxiliary busses, analyses along lines similar to those of Mr. Huening indicate that this margin disappears and the "tested capability according to the proposed rating basis" matches the more rigorous calculation almost exactly in the entire region from 1/2 cycle to 4 cycles after the start of short circuit.

Admittedly, the end-of-feeder example, Fig. 11, shows enough margin so that the more rigorous approach will sometimes be justified. However, it must be remembered that the duty at the end-of-feeder is considerably less than at the main bus. From the standpoint of future flexibility, it generally makes sense to use the same size breakers at both locations. Probably Mr. Huening has all this in mind when he states that his method "may supplement, though it is not intended to replace, simplified application calculations."

It is timely to call attention to the dangers of neglecting the induction motor contribution at contact parting (the basis for present standard application procedure). Relay

delay and breaker contact parting time are properly getting less and less attention. Whether or no the proposed rating basis is adopted, suitable corrective steps must be taken as soon as practical.

R. A. Huse (Public Service Electric and Gas Company, Newark, N. J.): This paper is of considerable interest to us in the utility field because we are confronted with the problem of holding auxiliary bus short-circuit duties within reasonable limits and at the same time providing adequate bus voltage when starting large motors. Under these conditions, with large induction motor loads, the motor contribution may be an appreciable part of the circuit-breaker duty.

In view of the increasing severity of this problem we have made several tests on auxiliary motors and busses to determine the extent of the induction motor contribution. Single motors as well as groups of motors were run at no load, then disconnected from their source and transferred to a bolted 3-phase short circuit in approximately 15 cycles. Representative oscillograms are shown in Figs. 14, 15, and 16. Current contributions have been corrected for the slight decrease in bus voltage during the de-energized period.

In using the equations given in Appendix I of the paper we suggest that the motor time constants be checked since we noted a large discrepancy between manufacturer's calculated time constants and those observed from test.

With 5-kv magnetic-type air circuit
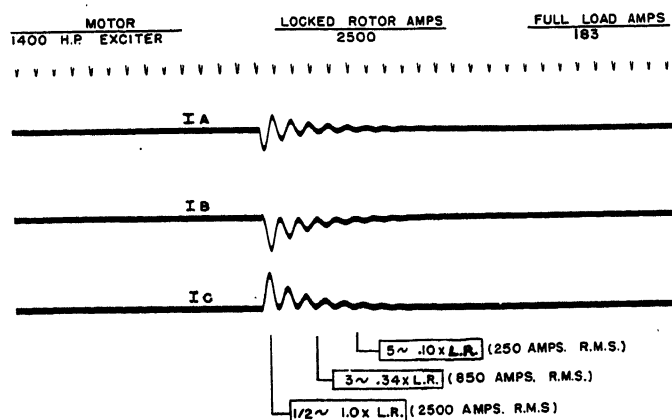
Fig. 14. Motor contribution to short circuits, test no. 1
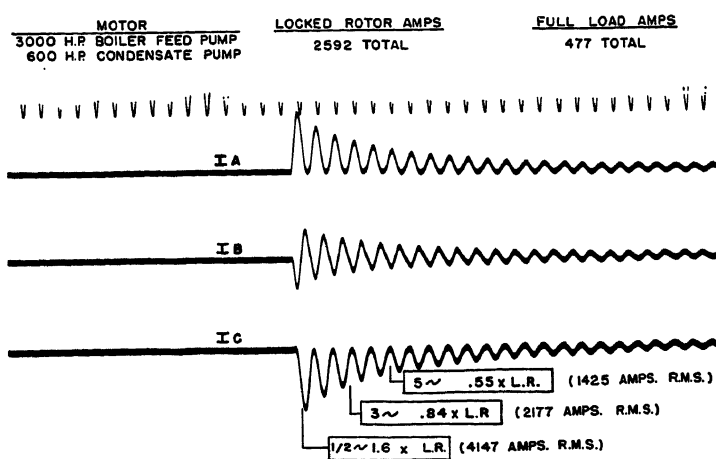
Fig. 15. Motor contribution to short circuits, test no. 2
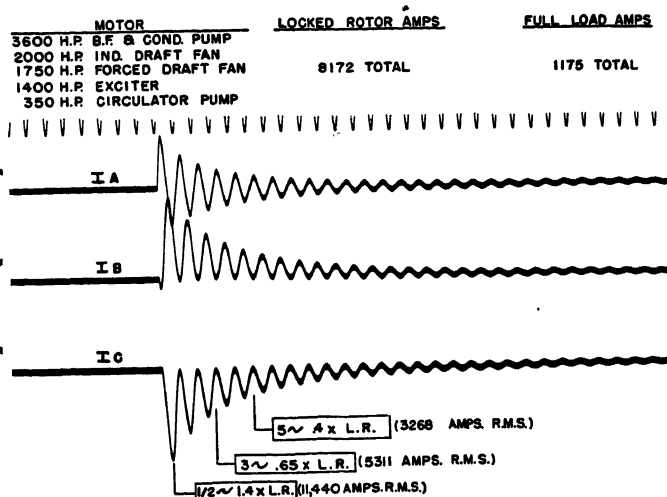
Fig. 16. Motor contribution to short circuits, test no. 6

breakers such as we use, which part contacts in approximately 3 cycles, our tests indicate that the induction motor contrbution is somewhat less than that suggested by the proposed new method of rating power circuit breakers. In view of this, if conditions are critical, it would be desirable to use Mr. Huening's methods which would give more accurate results.

Lawrence E. Fisher (General Electric Company, Plainville, Conn.): Mr. Huening's paper is a valuable contribution to the literature on this subject. It indicates that the feedback current from motors has a greater duration than was generally considered possible. To supplement the author's conclusions, I recall an experience in 1944 that tended to point toward the same conclusions.

An arcing fault was accidentally established in a 140-foot run of busway from a transformer bank to the switchboard in an industrial plant in which there were a great number of small motors. The arcing started near the input end of the busway which consisted of bare bus bars in a steel enclosure. This busway had a clear space inside the enclosure running the full length of the busway. By this is meant that there were no solid insulators serving as barriers to prevent the free movement of arcs. In such a busway any arc, even a single-phase arc, accidentally established within the enclosure will develop almost instantly into a 3-phase arc that will be driven by electromagnetic force almost like a shotgun blast in a direction away from the source of power.

Thus, previous experience had indicated that in this case the arc should have been driven to the switchboard end where it would have been confined by an ebonized asbestos barrier enclosing the end of the busway. Actually, however, and contrary to previous experience, the arc was driven at ultrahigh speed for 75 feet doing no damage traveling and then for no apparent reason, and with no physical barrier to even slow it down, the arc stopped and remained stationary long enough to completely burn off 1/4-inch by 6-inch copper bars. It was determined that the circuit breaker tripped it about 60 cycles.

Seeking an explanation for what I had considered an impossible behavior the short-circuit current was calculated. It was surprising to learn that the calculated short-

circuit current coming from the primary through the transformer bank was only about 20,000 amperes and the feedback current from the hundreds of motors in the plant was almost 15,000 amperes. This, of course, furnished a satisfactory explanation for the arc stopping where it did. The current from the transformer would force the arc toward the switchboard, but the feedback current would force the arc in the opposite direction. Therefore, the arc would move away from the source having the greater current value and in so doing would add busway impedance between itself and the greater source, at the same time subtracting busway impedance from the smaller source until it reached a point of equalization where the two currents would be about equal. Then the arc would stop and continue to burn copper until the circuit breaker tripped (in 1 second) and until the motor contribution died down. A labora-

tory test setup to simulate these conditions proved that the arc would always stop at the point of equalization of impedances between two equal sources.

This was the explanation given in my report to the customer indicating why the arc stopped where it did, although other conclusions and recommendations were, of course, included relative to the protective device characteristics required to prevent this type of trouble in the future.

This practical experience is given to supplement Mr. Huening's mathematical calculations which prove that the feedback current duration can be longer than generally thought possible.

D. G. Lewis (General Electric Company, Schenectady, N. Y.): I believe it is of interest to compare the results for one of Mr.

Fig. 17. Short-circuit currents and current magnitudes

Dashed lines show results of "exact" calculation short-circuit current at end-of-feeder of example

TIME IN CYCLES (OF A 60 CPS WAVE)
AFTER START OF SHORT CIRCUIT
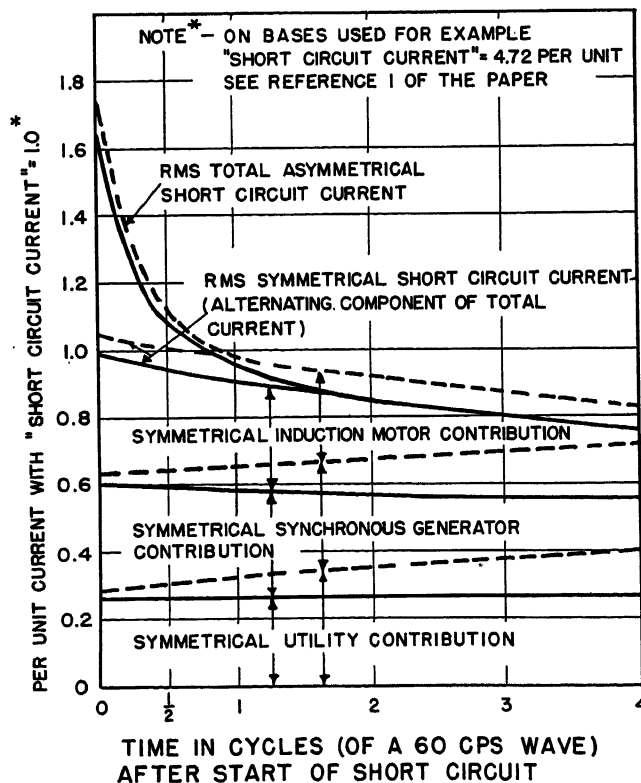
Huening's cases with results calculated in such a way as to account correctly for the common impedance in the fault branch. Consider the circuit shown in Fig. 4(A) and replace the synchronous machines and the induction machines by their operational impedances as shown at a and b.

(a): $x_d(p) = x_{LS} +$
$$\cfrac{1}{\cfrac{1}{x_{AD}} + \cfrac{1}{x_{FD} + \cfrac{R_{FD}}{p}} + \cfrac{1}{x_{KD} + \cfrac{R_{KD}}{p}}}$$

(b): $x_m(p) = x_{LS} + \cfrac{1}{\cfrac{1}{x_m} + \cfrac{1}{x_R + \cfrac{R_R}{p}}}$

These operational impedances give the correct short-circuit currents for the individual machines. The short-circuit currents for the circuit of Fig. 4(A) calculated using these operational impedances are shown as dashed lines in Fig. 17. As Mr. Huening points out, the utility contribution actually increases as the induction motor contribution decays.

There is an appreciable difference in the initial current magnitudes between these results and those obtained by the method of the paper (solid lines of Fig. 17). The reason for this is that the calculation that we have made based on operational impedances assumes that the initial symmetrical current magnitudes are limited by reactance only. For machines alone this is a good assumption, but when a feeder is included with relatively high resistance, as in this case, the value of the resistance also affects the initial symmetrical current magnitudes. For this reason, Mr. Huening's results are more nearly correct, in the first cycle, than

ours. Either the results of the operational solution can be normalized to give the correct initial value, in which case the result for total symmetrical fault current will agree very closely, or more detailed calculations can be made to include the effect of stator-circuit resistances.

The direct current can also be calculated operationally and combined with the symmetrical component to give total rms current. The results for this current also agree very well if the currents derived from the operational solution are normalized to give the correct initial value.

W. C. Huening, Jr.: Mr. Webb indicates that the method of calculation described in the paper has already proved helpful to a special Working Group set up by the Committee on Switchgear. It is indeed gratifying to learn that the method is being applied by others with useful results.

Mr. Woodrow, as Chairman of the Working Group on Methods of Rating, made the inquiry about motor contributions which led to an investigation and resulted in the preparation of this paper. The discounting of induction motor contributions mentioned by Mr. Woodrow is apparently a satisfactory method of including in a simplified calculation the effects of motor short-circuit current variations without placing undue emphasis on them. As Mr. Woodrow points out, only occasionally will there be reason to use a longer, more rigorous calculation, such as the one described in this paper. The simple calculation proposed by the Working Group in reference 1 will almost always select an adequate circuit breaker usually without excessive margin. However, when it is felt that some margin exists that will never be needed, the method of calculation from the present paper may

be useful in estimating the amount of this margin.

Mr. Huse supplies some results of field tests on an installed induction machine and compares the test results with calculations. The conclusions Mr. Huse draws from his comparison should prove valuable. His suggestion that time constant information be checked is particularly worth while for instances when contributions from large induction motors close by are important portions of a total short-circuit current, because accurate data aid in obtaining accurate results. Note that groups of smaller motors, especially those contributing through circuit impedance, may be important in other cases. The only feasible way of evaluating the contribution from these groups is with approximations, and the results obtained are usually satisfactory. It is also interesting to note that Mr. Huse finds the proposed new method of rating power circuit breakers conservatively evaluates the short-circuit contributions from induction motors close by.

Mr. Fisher provides an interesting description and explanation of the behavior of an arcing short circuit in a busway system. His explanation tends to further support the conclusion that motor contributions persist for longer than is usually assumed.

Mr. Lewis gives the results of even more accurate calculations for one of the examples from the paper, and compares the two sets of results. The close agreement is support for the statement that the method proposed in this paper gives good results up to 4 or 5 cycles after the start of the short circuit.

I wish to take this opportunity to express my thanks for the contribution made by all five discussers. Their comments and invaluable additional material are much appreciated.

Underground Corrosion on Rural Electric Distribution Lines

O. W. ZASTROW
MEMBER AIEE

SOME studies of underground corrosion associated with the grounding of multigrounded wye-connected distribution lines are described in this paper. Underground corrosion damage, mainly of anchor rods, has been experienced in enough locations to deserve special attention, but there seems to be very little electric utility experience to draw on in deciding what to do about it. The lack of information on this subject in the literature may be explained by reasons such as the following:

1. The widespread adoption of multigrounded wye-connected distribution lines is fairly recent, having occurred mostly

in the last 15 to 20 years. Compared with delta-connected lines, multigrounded lines have a greater multiplicity of ground connections and therefore offer more possibilities for undesirable corrosion reactions.

2. Some of the most severe conditions causing underground corrosion are in areas where relatively little mileage of distribution line existed prior to 1935 or 1940. As a result, comparatively little experience in dealing with these conditions has been accumulated.

3. Judging from literature on this subject (or the lack of it), the electric utility industry has not given much attention to the effects of electrical grounding on underground corrosion.

This summary of experience to date is

offered for the benefit of those who may encounter similar problems. It is hoped that others who can contribute to a better understanding of the factors involved will add their comments and suggestions.

How Corrosion Occurs

ELECTROCHEMICAL NATURE OF CORROSION

Corrosion is generally regarded as an electrochemical process in which the chemical reactions are associated with electrical effects. This concept is most helpful for considering corrosion in relation to grounding and therefore will be discussed in some detail.[1]

The electrochemical concept of corro-
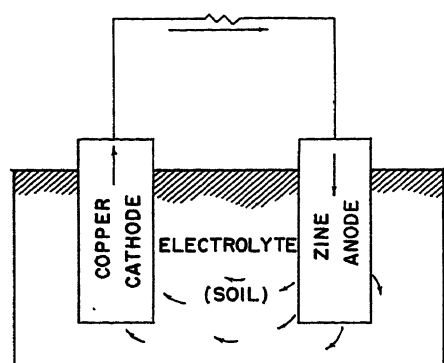
Fig. 1. Galvanic cell. Arrows show direction of current flow

sion should seem entirely natural to any electrical engineer if some elementary rules of chemistry and physics are kept in mind. For example, a metal usually dissolves in its environment (corrodes) by forming positively charged ions that leave the metal surface. In other words, corrosion is associated with a flow of direct current from the metal to the surrounding medium. Also, a metal that is relatively active chemically loses more positive ions and therefore assumes a more negative d-c potential than a metal that is less active. This is illustrated by Table I, which shows probable equilibrium potentials of several metals in earth, using copper as a reference. The open-circuit d-c potential of a metal is generally more negative in a corrosive soil than in a relatively noncorrosive soil, since the metal tends to discharge ions more vigorously in the more corrosive soil.

An electrochemical process is illustrated by the galvanic cell of Fig. 1. If copper and zinc electrodes in conducting soil are connected together, a potential of 1.1 volts tends to cause a current flow as shown. Anodic reactions at the zinc electrode cause the zinc to corrode, while the cathodic reactions at the copper electrode protect it against corrosion. The amount of current flowing is limited by the resistance of the electrolyte, by the resistance of the external circuit and by various polarization effects at or near the electrode surfaces.

Table I. Open-Circuit Potential in Earth Volts (to a Copper Electrode)

Metals	Equilibrium Potentials
Magnesium	−2.7
Aluminum	−2.0
Zinc	−1.1
Iron	−0.7
Lead	−0.4
Copper	0

GROUND CONNECTIONS ON MULTIGROUNDED LINES

Rural Electrification Administration type distribution lines have a primary neutral conductor that is common with the secondary neutral at transformers and other locations where secondary or service conductors exist. The neutral is grounded in accordance with the requirements of Paragraph 97C1 of the National Electrical Safety Code[2] and Article 250 of the National Electrical Code.[3] Additional grounds are usually provided by anchor rods and anchors because, in most cases, the guys are connected to the neutral and are not insulated. This is permitted under the conditions set forth in Paragraph 283B4 of the National Electrical Safety Code.[2] Pole protection grounds, where used, are also connected to the system neutral. The driven primary grounds and pole protection grounds are generally copper-covered or solid copper. The nominal operating voltage on primary distribution lines is 7.2/12.47 kv in most areas and 14.4/24.9 kv in some areas of lower consumer density.

With the connections just described, the neutral conductor provides a continuous metallic circuit between the following buried structures:

1. Driven ground electrodes, usually copper-covered 5/8-inch by 8-foot rods, used at all transformers, at intervals of not more than 1,500 feet along primary lines, and at services. Driven pipe electrodes of galvanized steel are used at services in some cases.

2. Pole protection ground electrodes, "butt" type, terminated at a copper plate or coiled copper wire, no. 6 American Wire Gauge, at the base of the pole. These are used where lightning conditions are severe, and are frequently omitted where lightning is less severe.

3. Galvanized steel anchor rods, usually 5/8 inch in diameter by 7 or 8 feet long, and steel anchors.

4. Grounds on consumers' premises, including buried metal pipes and well casings as well as the frames of machines associated with electric equipment. These may be copper, bare or galvanized steel, or combinations of these metals.

5. Buried metal conduit occasionally used at services, on consumers' premises, and at station metering installations.

6. Substation or generating station structures and grounds, including grounding mats and ground connections to static wires where present and interconnected.

The multigrounded wye-connected distribution line has proved to be highly satisfactory from the standpoint of electrical performance, including safety, and from the standpoint of over-all economy. It also provides conditions that can be
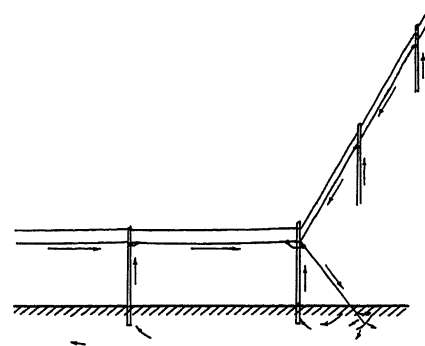
Fig. 2. D-c flow due to galvanic action between copper grounds and a steel anchor rod

favorable to underground corrosion, as will be described in more detail.

GALVANIC CORROSION

Referring to Table I, a steel electrode in earth tends to assume a potential approximately 0.7 volt negative with respect to a copper electrode similarly buried. If the steel is galvanized, the potential difference is approximately 1.1 volts. If copper-covered ground rods and steel anchor rods are connected to a common neutral conductor, the copper versus steel potential tends to cause a flow of current as shown in Fig. 2. This current is associated with galvanic corrosion at the steel (anodic) surfaces. Effects at the copper surface tend to protect it against corrosion and may precipitate soil salts, such as calcium carbonate, as a result of alkaline conditions caused by the cathodic reaction.

Galvanic corrosion does not require the presence of dissimilar metals. Anodic and cathodic areas may form on a metal surface because of variations in the structure of the metal or its environment. For example, an anchor rod and anchor buried in earth may be exposed to varying concentrations of oxygen as shown in Fig. 3. Other conditions being equal, the iron surface tends to become anodic where oxygen is excluded and cathodic where oxygen is present, with the result that galvanic corrosion removes iron from the most deeply buried surfaces.[1]
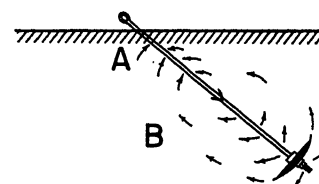
Fig. 3. Galvanic corrosion of an electrically isolated steel anchor rod and anchor

A—Cathodic area where oxygen is present
B—Anodic area subject to corrosion where oxygen is excluded
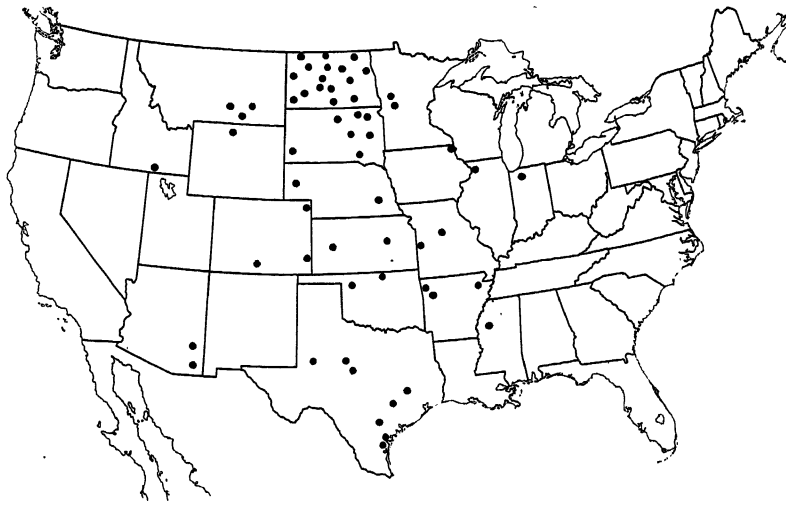
Fig. 4. Locations of electric distribution systems experiencing anchor rod failures or excessive corrosion 15 years or less after original construction

Stray Current Corrosion

Direct currents from external sources sometimes flow in grounded metal structures and cause corrosion where the current flows from the metal into the soil. Cathodic protection systems for underground pipe lines may cause such corrosion of ground electrodes, anchor rods, or consumer plumbing on rural distribution lines under some conditions.

Scope of Problem

No Difficulties Are Reported in Many Areas

The foregoing discussion and Fig. 2 suggest that excessive corrosion is likely to occur wherever steel anchor rods (or other buried steel structures) are connected to the neutral of a copper grounded distribution line. Fortunately, that is not the case. Experience in many areas of the United States has revealed no difficulties from this source during the 17
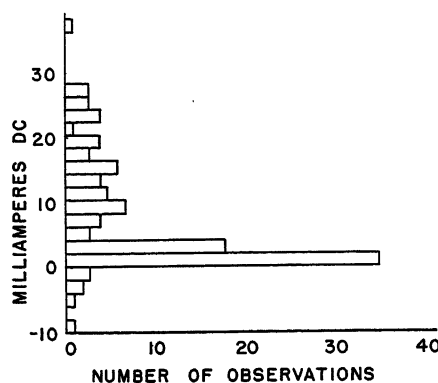
Fig. 5. Distribution of galvanic currents measured in guys. Positive values indicate current flowing toward anchor rods. Total number of observations, 108

years or more that Rural Electrification Administration-financed lines have been in operation. Apparently, polarization and soil resistance reduce the galvanic currents to such an extent that the corrosion damage is not serious in a majority of soils.

Corrosion, Where Experienced, Is Variable

Corrosion damage to anchor rods has been reported by approximately 60 rural electric distribution systems located as shown in Fig. 4. (The total number of rural distribution borrowers is 938 as of December 1953.) The damage is noticed when lines are relocated or where one or more anchor rods part completely, allowing guys to "go slack." Failures of anchor rods occur within approximately 4 to 10 or 12 years in the worst locations, but the rate of corrosion is widely variable throughout each of the areas where conditions have been examined in detail. This is helpful from the standpoint that damage is generally noticed before any substantial percentage of the anchor rods is seriously weakened. However, determinations of the extent of damage in any particular case are made more difficult because of the variations in rates of corrosion.

The galvanic currents flowing in guys and other ground connections provide a helpful indication of the rate of corrosion, since the current is directly proportional to the amount of metal being removed at anodic surfaces. This is in accordance with theoretical considerations and has been verified experimentally.[1] Corrosion from local action (as shown in Fig. 3) is not revealed by measurements in an external connection such as a guy or ground wire. However, our experience

indicates that corrosion of anchor rod is generally associated with galvanic currents flowing in guys.

The variability of currents measured in guys is illustrated by Fig. 5, which is based on 108 observations at locations that were chosen without regard to the factors affecting corrosion rates. Fig. 6 shows similar data on the currents measured in the ground wires connecting copper-covered ground rods to the neutral.

Which Soils Are Most Corrosive?

The soils most likely to be corrosive can usually be identified from terrain features, characteristics of the soil, or both. Tests described by Denison and Romanoff[4,5] indicate that underground corrosion is most severe in poorly aerated, reducing soils that are highly acid or contain high concentrations of soluble salts. Such conditions occur in low or level locations that are poorly drained and result in low soil resistivity. Light-colored mottling or streaks in tight clay subsoils indicate a reducing, highly corrosive soil.

The corrosive soils observed during these tests are in areas of relatively low rainfall, as may be noted in Fig. 4, and are in poorly drained locations where soil resistivity is low. In North and South Dakota, grey to white mottling is clearly evident in corrosive subsoils and in some cases white alkali is visible on the surface. Earth resistivities to an average depth of 10 feet were generally less than 20 meter ohms (2,000 ohm-centimeters) in corrosive locations.

Components Affected by Corrosion

Where corrosion is excessive the damage is generally noticed on anchor rods first. This is true in cases of stray current corrosion as well as for galvanic corrosion damage. However, other buried steel connected to the neutral conductor may be similarly affected. Electrical measurements have indicated galvanic currents up to 200 milliamperes (ma) or more discharged by buried pipes, well
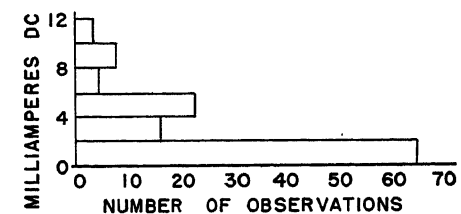
Fig. 6. Distribution of galvanic currents measured in connections to copper-covered ground rods. Positive values indicate current flowing from ground rods toward neutral. Total number of observations, 117
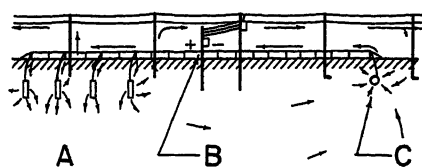
Fig. 7. Diagram of pipe-line protection system served from a multigrounded distribution line. Arrows indicate direct currents

A—Anodes
B—Rectifier
C—Protected pipe

casings, underground steel conduit, and, in one case, a cotton gin. These cases have been infrequent, but the damage that may result is enough to justify special precautions against any such occurrences.

STRAY CURRENT CORROSION

A few cases have been found where pipe-line protection systems caused corrosion of distribution line grounds and, in one case, of consumer plumbing. In each case, the rectifier received energy from the distribution line affected and electrical ground connections were located near to anodes of the cathodic protection system. The components involved and the flow of direct currents that may result are shown in Fig. 7. It may be noted that most of the damage occurs at some distance from the rectifier, where the currents eventually must return to earth.

A serious situation may result if a consumer is served from a multigrounded electric distribution line and also from a cathodically protected gas distribution line. In such cases, the multigrounded primary neutral conductor is connected to the secondary neutral and thence to any water pipes and wells on the consumer's premises. If a connection to the protected gas pipe is made, either deliberately or inadvertently, cathodic protection to the pipe line may be adversely affected. If no connection to the protected gas pipe line is made, stray currents may cause rapid corrosion of the consumer plumbing. The only safe procedure is to examine each installation carefully and
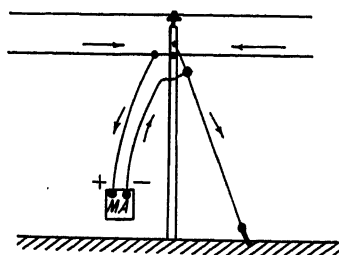
Fig. 8. Measurement of galvanic current in a guy using a d-c milliameter

apply the necessary measures (i.e., physical co-ordination supplemented by resistance bonding connections), preferably at the time the electric service or gas line is installed. The likelihood of serious damage is greatly reduced if the gas service pipe is insulated from the main, as is the general practice in many areas.

Analysis and Survey Methods

The approach to underground corrosion on rural distribution lines is influenced by two considerations that may seem unusual, in degree at least, to most corrosion engineers:

1. Individuals with no previous specialized experience in corrosion mitigation will generally make the determinations as to whether excessive corrosion is occurring and, if so, what should be done about it. This seems applicable to system engineers (whether retained or employed full time) as well as to other employees, since engineering services have been concerned mainly with electrical and mechanical design.

2. The average investment and the annual operating expense is relatively low in terms of dollars per mile of line, total costs notwithstanding; so measures taken against underground corrosion need to fit into the same pattern. This requires a minimum of special designs for individual locations and a minimum of follow-up checking and maintenance to assure reasonable freedom from excessive underground corrosion in the future.

DETERMINING LOCATIONS AND
CORROSION RATES

To the distribution system operator concerned about underground corrosion, the most difficult problem is to obtain a reliable estimate of the rate of damage at various locations. The importance of such information can hardly be overemphasized, since the choice of mitigative measures is generally based on assumptions, right or wrong, as to corrosion rates and the locations affected. Efficient application of measures chosen requires the best possible understanding of the existing conditions, from the standpoint of effectiveness and economy and also in the interest of applying the measures first where they are most needed. Experience has shown that the first evidence of excessive underground corrosion is noticed on anchor rods, if the steel anchor rods are connected to the common neutral along with copper-covered ground electrodes.

VISUAL EVIDENCE

Underground corrosion usually is not visible, yet visual checks can provide the
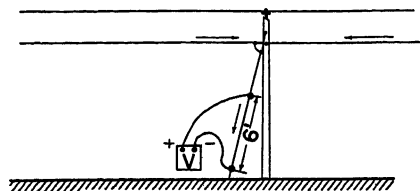
Fig. 9. Determination of galvanic current from voltage drop along a guy

alert operator with much valuable information. For example, anchor assemblies must frequently be removed because of road widening operations or for other reasons. The condition of anchor rods removed from service gives valuable evidence as to whether or not underground corrosion is occurring at an excessive rate. If corrosion is evident, the condition of the anchor rod and its age give an indication of how long it might have lasted before becoming seriously weakened.

The number of changes the distribution system operator is plagued with will usually provide numerous opportunities for visual examinations of anchor rods that have been in service for varying lengths of time. By observing the condition, age, and location of each anchor rod, he should be able to determine whether or not a corrosion problem exists and, if so, some of the troublesome locations (and trouble-free locations). The rate of damage at the worst locations will help to establish the probable length of time before failures may occur.

Data on anchor rods that have been removed from service or replaced because of failure due to corrosion provide the only direct information on rates of corrosion. Such information should be carefully noted at every opportunity if corrosion problems are suspected. Visual checking of anchor rods in place requires practically complete removal of each rod, which is costly and time-consuming.

ANCHOR ROD CORROSION AND GALVANIC
CURRENTS IN GUYS

Underground corrosion, when serious enough to deserve special attention, is generally associated with galvanic currents that can be measured above ground. Some exceptions to this may occur (see Fig. 3, e.g.), but such have not yet been encountered in our experience.

Galvanic current measurements in guys have proved to be the only practical means for estimating the rates of damage to individual anchor rods if extensive surveys are desired. The indications are only approximate, but this does not prevent them from being useful if the limitations are recognized. The two most important limitations are:
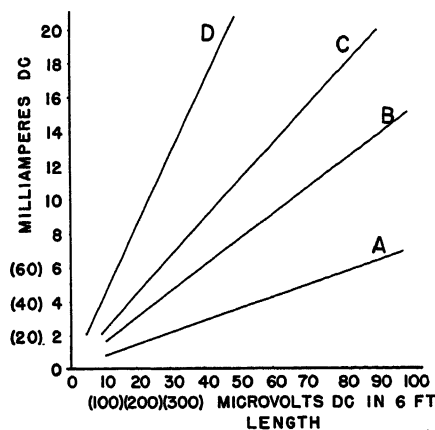
Fig. 10. Sample calibrations for voltage drop determination of galvanic current, for common sizes of Diemens-Martin guy strand and ground wire

A—1/4-inch guy strand
B—3/8-inch guy strand
C—7/16-inch guy strand
D—No. 6 American Wire Gauge soft drawn copper

1. The galvanic reaction and the current are subject to variations due to changes in soil moisture and temperature. As a result, absolute values of current at a particular time mean little in terms of metal removed until correlated with visual observations.

2. The amount of metal that can be lost before an anchor rod is seriously weakened varies, depending on whether the damage is concentrated or spread along most of the buried portion.

A correlation between galvanic currents in guys and rates of damage to anchor rods can be established by measuring currents and then removing the anchor rods for visual examination. This is time-consuming and relatively expensive unless it can be done when the removal of anchors is necessary for other reasons.

If anchor rod failures or specified amounts of damage have resulted after known lengths of time, measurements of galvanic currents in guys may help to show the scope of the damage even if the exact locations of the rods affected are not known. For example, Fig. 5 shows the distribution of currents in guys observed in July 1954 on the lines of the Douglas Electric Co-operative, Inc., Armour, S. Dak. The first failure of an anchor rod on this sytsem occurred after 5 years of operation. Other anchor rods removed because of road widening were corroded to various degrees at the threaded (bottom) ends, so that the ones most affected might have failed in another year or two. These visual observations, together with the currents measured, may suggest steps such as the following:

1. Where the product of d-c ma toward ground by years in service is 100 or more (20 ma or more in a 5-year-old guy), replace all critical anchor rods and examine the rods removed for additional data, which will provide guidance as to whether the others in this group should be replaced.

2. For current time products of 50 to 100 ma years (10 to 20 ma for a 5-year-old rod), replace a few of the most important anchor rods to obtain additional visual information. Many anchor rods in this group may still be serviceable.

3. Wherever the current is approximately 4 ma or more, use appropriate measures to prevent further flow of direct current or to reduce it to less than 4 ma.

These ma figures are given to illustrate a method of approach and are subject to variations depending on soil moisture, season, and terrain. Observations made near Corpus Christi, Tex., in August 1951 indicated anchor rod failures after only 50 ma years of current flow. At the other extreme, an anchor rod near Devils Lake, N. Dak., was discharging a current of approximately 45 ma in June 1953 and was still holding after nearly 5 years in service. The rod was removed and it was found that the damage was very evenly distributed, with the cross section reduced to approximately one-fourth the original area at the weakest point, a few inches above the anchor.

The galvanic current in a guy can be measured with a d-c milliameter as shown in Fig. 8, and some general-purpose meters of types used for radio testing can be used. The measurement becomes relatively inaccurate if resistance of the meter is greater than the resistance to ground of the anchor rod, and a correction for meter resistance is needed unless resistance of the meter is approximately 1 ohm or less.

The method illustrated in Fig. 8 has been found to be too time-consuming where many observations are needed, since it is necessary to climb a pole and disconnect the guy from the neutral each time a measurement is made. As a result, most of the current determinations during this study have been made from measurements of direct voltage drop along guys and ground wires; see Fig. 9. The voltages to be measured are very small for currents of only a few ma, as may be seen from Fig. 10, so that a very sensitive instrument is needed and care is necessary to avoid errors resulting from contact potentials. The instrument used for the current determinations referred to is a Weston model 622 voltmeter with a 0 to 1-millivolt range and 100 scale divisions. Special vise-type pliers were found useful as test clamps to provide adequate contact pressure. The instrument and test leads used are shown in Fig. 11.

DIRECT CURRENTS IN NEUTRAL CONDUCTORS

Measurements of direct current along the neutral conductor have been used to determine whether power line grounds may be damaging a consumer's plumbing or vice versa. Most of these were made by a zero resistance ammeter method or by use of the improvised millivoltmeter shunt shown in Fig. 12.

POTENTIAL MEASUREMENTS

Measurements of the d-c neutral potential were made frequently during this study. It was hoped that some correlation might be found between d-c potential
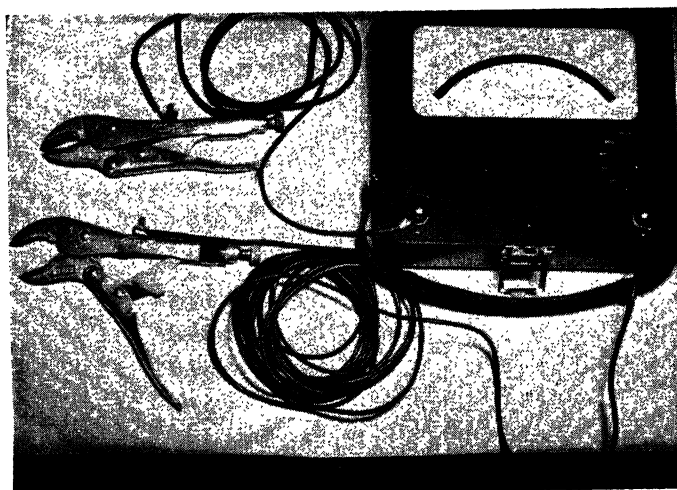
Fig. 11 (left). Millivoltmeter used for voltage drop determinations of galvanic current

Fig. 12 (right). Millivoltmeter shunt for estimating direct current to or from a service neutral. For no. 12 American Wire Gauge solid copper wire, the voltage drop is approximately 10 microvolts per ma for length L=6.5 feet
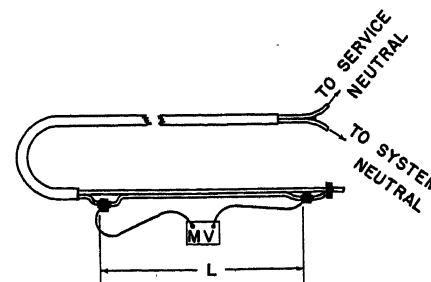
and corrosion conditions, since the potential could be readily measured without climbing poles.

The d-c potential measurements have been found too difficult to interpret in terms of probable corrosion damage to take the place of current measurements for survey purposes, since a given potential may result from several different combinations of conditions. Potentials observed have generally varied from −0.35 to −0.75 volt, except for a few cases where more extreme potentials were caused by pipe-line protection systems. All potentials were measured from a copper-copper sulfate reference electrode.

Mitigative Measures

Requirements to Be Considered

Corrosion can be prevented in practically any situation if cost may be disregarded. The problem of the distribution system operator is to decide what measures will assure adequate performance of distribution line components at the least cost consistent with safety and consumer satisfaction. The measures selected should:

1. Reduce rates of corrosion so that the life of underground components of the line will be extended to at least 25 to 35 years, the figure depending on circumstances.

2. Minimize the possibility of any adverse effects to consumer plumbing or other buried metal connected to the system neutral but not part of the distribution line, due to interconnection with the distribution line neutral. This may not be done at the expense of the consumer grounding specified in Paragraphs 2581 and 2582 of the National Electrical Code.[3]

3. Provide, as far as possible, for uniform construction specifications throughout the distribution system.

4. Keep to a minimum the number of locations where special measures are needed, particularly if special follow-up or maintenance measures are indicated.

5. Comply with requirements of the National Electrical Code, the National Electrical Safety Code,[2] and additional state and local regulations where applicable.

In addition to these requirements, consideration is necessarily given to economy and to any possible adverse effects on operating characteristics of the distribution system.

The Measures that Seem Most Useful

Three general types of mitigative measures appear to be useful and practical for relieving corrosion damage to anchor rods or other buried steel affected by galvanic corrosion. They are:

1. Insulate buried steel from the copper-

grounded neutral or replace the steel components with other materials that will not corrode excessively when connected to the copper grounded neutral.

2. Replace copper ground electrodes with galvanized steel ground rods.

3. Install sacrificial anodes to provide drainage and to prevent damage to the anchor rods or other buried structures.

Each of these measures has advantages and limitations that change in relative importance with local conditions.

The Copper Grounded Neutral Approach

Strain insulators in guys and corrosion-resistant anchor rods (bronze, copper-covered, or stainless steel) will be considered together, because either results in an approach to an all-copper grounded neutral, in so far as galvanic effects are concerned. Guy strain insulators ordinarily will be less costly and therefore are a more likely choice unless it is believed that a galvanized steel anchor rod will corrode at an excessive rate even if insulated from the neutral. Corrosion-resistant anchor rods probably should be used only with nonferrous (concrete or creosoted log) anchors; otherwise excessive galvanic corrosion of the anchors may result.

A strain insulator or corrosion-resistant anchor rod and anchor offers the simplest remedy where a particular anchor rod is subject to excessive corrosion, and it may be the least costly. The most important disadvantage is that the damage may be shifted to near-by anchor rods or to other buried steel still connected to the neutral, unless additional steps are taken to prevent it. Therefore the following limitations should be considered:

1. Under some soil conditions, strain insulators or special anchor rods and anchors, if used, may have to be applied to all anchor assemblies to prevent accelerated damage to steel anchor rods still connected to the neutral.

2. Special measures are needed wherever buried plumbing, conduit, or other structures of steel are connected to the neutral. If any of these are overlooked, relatively rapid corrosion may result.

Corrosion resistance of the copper-covered ground rods and anchor rods now available may not be adequate in some corrosive soils on an all-copper grounded system, judging from test results reported by Denison and Romanoff.[4] The test data show that after 14.3 years the penetration of tough pitch copper exceeded 0.010 inch at 7 of the 14 test sites and 0.020 inch at two test sites. The minimum copper thicknesses on 5/8 inch copper-covered ground rods available in the United States is generally between

0.010 and 0.020 inch. Experience with the copper-covered ground rods apparently is satisfactory, but the experience may not be applicable to all soils now being encountered. Also, copper ground electrodes and other structures receive cathodic protection at the expense of buried steel and other metals in many cases, and this tends to prolong the life of the copper underground.

Steel Grounding

Galvanized steel ground rods, used instead of copper or copper-covered grounding electrodes, reduce galvanic corrosion of other buried steel connected to the system neutral. The galvanic current flowing toward the neutral is reduced as the amount of buried copper is reduced and, in addition, the steel ground electrodes discharge part of the current that would otherwise return to earth through guys and anchors.

The steel grounding approach is particularly useful in areas where corrosion is relatively slow and general over a considerable area, since a general reduction in corrosion rates can be achieved by replacing some ground electrodes on an existing line without the necessity for changing all grounds.

The most important limitations where galvanized steel ground rods are used are:

1. There does not appear to be a steel equivalent at present to the copper "butt wrap" type of pole protection ground, particularly for use on lines that have some copper grounding electrodes, so that at present only driven steel electrodes can be used.

2. A copper to steel connection underground is necessary, since copper ground wires are used. The copper wires are considered necessary for adequate conductivity at most locations and appear to be desirable at other locations (as at pole protection grounds) for corrosion resistance at the ground line.

If steel ground rods are used, damage to grounds and anchor rods may still occur at locations where consumers have copper underground plumbing or where the distribution neutral is interconnected with copper grounds at the power source. Special measures will sometimes be needed at such locations.

Galvanized steel ground rods considered for use at present are 3/4 inch in diameter and hot-dip galvanized in accordance with Specification A153-49 of the American Society for Testing Materials,[6] to an average coating thickness of 2 ounces per square foot. The 3/4-inch diameter may seem inconsistent with the requirement for anchor rods, which are 5/8 inch in diameter and similarly galva-
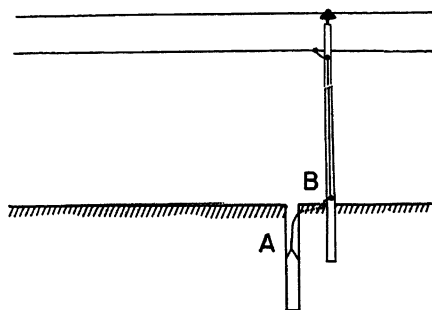
Fig. 13. Method of installing a galvanic anode

A—Anode with backfill material and connecting lead

B—Connection to ground wire for previously disconnected pole protection ground

nized. However, the larger diameter ground rod is regarded as necessary at present for compliance with Paragraph 95D of the National Electrical Safety Code.[2]

The galvanized steel ground rods are regarded as entirely adequate in resistance to corrosion. This conclusion is based on more than 15 years' experience with anchor rods installed under the adverse conditions that have been discussed and on test results reported by Denison and Romanoff.[5]

Cathodic Protection by Sacrificial Anodes

Anodes of magnesium, zinc, or aluminum protect underground metal against corrosion by providing drainage points where direct current is discharged into the earth. This causes the d-c potential of the protected structure to become more negative and, if carried far enough, causes all buried parts of the structure to become cathodic (with current flowing toward metal from the soil at all surfaces) so that complete protection is achieved.[1]

Sacrificial anodes can be readily installed on a multigrounded distribution line as shown in Fig. 13. Anodes with backfill material included and the connecting leads attached (such as the "Galvo-pak" magnesium anodes) are desirable when installations are to be made by inexperienced personnel. The anodes should be installed at the lowest locations available, at depths at least as great as the metals being protected. The choice of anode material is influenced by earth resistivity, current requirements, the length of life desired before replacement becomes necessary, and cost. Magnesium produces the highest current output for a given anode size and soil resistivity, since it has the highest driving voltage. Zinc has the advantage of higher efficiency at relatively low out-

puts. Relatively little data on aluminum anodes are available at present, but the solution potentials measured suggest that the driving voltage is approximately the same as for zinc.[7]

Sacrificial anodes are the most useful at special locations where strain insulators or steel grounding alone will not prevent excessive corrosion. For example, on a copper grounded line the anodes may be used at locations where buried steel cannot be disconnected from the system neutral. On steel grounded lines, sacrificial anodes are useful at substations where the system neutral is connected to copper station or transmission-line grounds, or at locations where underground copper plumbing or other structures are connected to the neutral.

Sacrificial anodes gradually expend themselves so that replacement becomes necessary. For example, a 17-pound magnesium anode lasts approximately 10 years at a current output of 100 ma or proportionately less at higher currents. Weight losses at various current outputs have been determined in tests, and they can be calculated from the handbook values of electrochemical equivalents, for any assumed efficiency and valence of the anodic reaction.[7,8]

Some field trials of sacrificial anodes connected to a distribution line neutral have been undertaken for the purpose of determining what current output may result and the range of protection that may be achieved. Unfortunately, results of the trials are not available at the time of this writing. The types of anodes included are:

1. Type-*17D* (17-pound) "Galvo-pak" magnesium anodes, which include backfill material.

2. American Zinc Institute type-*AZI-2-30* (30-pound) zinc anodes with backfill of gypsum plaster and native earth.

3. Aluminum Company of America style-*D 109* (3-inch diameter by 30-inches long) aluminum alloy anodes in backfill as supplied by the manufacturer.

All anodes are installed at depths of 8 to 10 feet at locations where the average earth resistivities to a 10-foot depth are between 4 and 10 meter ohms (400 to 1,000 ohm-centimeters) as measured in July 1954. Resistance to ground of the system neutral at points where the anodes are connected is 0.7 to 1.3 ohms.

STRAY CURRENT CORROSION

Damage due to pipe-line protection rectifiers as illustrated in Fig. 7 can be prevented by the measures shown in Fig. 14, as follows:

1. Remove the pole protection grounds *A*

that are less than 200 feet from buried anodes, to prevent direct current from flowing onto the system neutral. If any anchors are installed here, install strain insulators in the guys.

2. Request a resistance bond (installed by the pipe-line agency) at *B* to drain from the neutral at least as much current as flows toward the neutral at any ground connections that must be retained for safety (as at *C* and *D*) near the buried anodes. The resistance bond is normally inside the rectifier cabinet, and some pipe-line agencies apparently make a practice of including such bonds in all rectifier installations.

3. Remove any nonessential ground (such as *E*) located so near the pipe line that damage cannot be prevented. If an anchor is present at such a location, install a strain insulator in the guy.

From a practical standpoint, the simplest method of avoiding damage in most cases is to make certain that the corrosion specialist responsible for the rectifier installation is advised of the power grounding arrangement used.

Some stray current problems may require special analysis and special mitigative measures. An example is the situation where a consumer with underground plumbing is served from a cathodically protected gas distribution line and also from a multigrounded electric distribution line. Such a situation requires cooperation between the pipe-line and electrification engineers concerned. Procedures for handling of joint cathodic protection problems are outlined in a report on cathodic protection.[9]

Conclusions

Additional study and experience are desirable on a number of aspects of the subject discussed. This is particularly true in regard to survey methods for determining the extent of corrosion on any particular distribution line and the choice and application of effective mitigative measures at minimum cost.

The studies and experience that have been discussed point to the following conclusions:

1. Excessive underground corrosion on multigrounded distribution lines is generally a result of galvanic action between buried copper and steel connected to the system neutral.
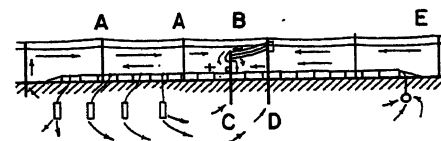
Fig. 14. D-c flow near a cathodic protection rectifier with power grounds completely protected

2. Stray current corrosion due to cathodic protection rectifiers may cause relatively rapid underground corrosion but should not present any serious problem if reasonable precautions are taken.

3. Galvanic corrosion of anchor rods on existing lines can be relieved by the use of strain insulators in guys or corrosion-resistant anchor rods, by substitution of galvanized steel ground rods for copper electrodes, or by the use of galvanic anodes. Each of these measures has advantages and limitations that change with local conditions. Steel grounding seems preferable in most cases from the standpoint of economy and minimum danger of corrosion damage to other buried structures.

4. Galvanized steel ground rods should generally be used for replacement and new construction on all lines constructed with steel anchor rods bonded to the neutral, to minimize the possibility of excessive underground corrosion. This appears applicable to multigrounded common neutral systems generally except where strain insulators or corrosion-resistant materials are being applied for corrosion prevention.

5. Anchor rod corrosion is the first indication of galvanic corrosion on multigrounded lines with anchor rods connected to the neutral, but this should be recognized as a symptom and not necessarily the entire problem. All underground structures connected to the common neutral conductor should be considered when planning or applying mitigative measures.

6. The attitude that if some grounding is good more is better needs to be reconsidered from the standpoint of underground corrosion. Grounding in excess of the requirements for safety and lightning protection may be a major factor contributing to excessive underground corrosion.

References

1. Corrosion Handbook (book), Herbert H. Uhlig. John Wiley and Sons, Inc., New York, N. Y., 1948.

2. Safety Rules for the Installation and Maintenance of Electric Supply and Communication Lines. National Bureau of Standards Handbook H32, Washington, D. C., Sept. 23, 1941.

3. National Electrical Code—Volume V. National Fire Protection Association, Boston, Mass., 1953.

4. Soil-Corrosion Studies, 1946 and 1948: Copper Alloys, Lead, and Zinc, Irving A. Denison, Melvin Romanoff. Research Paper RP2077, Journal of Research, National Bureau of Standards, Washington, D. C., vol. 44, Mar. 1950.

5. Corrosion of Galvanized Steel in Soils, Irving A. Denison, Melvin Romanoff. Research Paper 2366, Ibid., vol. 49, no. 5, Nov. 1952.

6. Specification A153–49, American Society for Testing Materials, Philadelphia, Pa., 1949.

7. First Interim Report on Galvanic Anode Test of Technical Practices Committee No. 2, Galvanic Anode Committee. Publication No. 50–2, National Association of Corrosion Engineers, Houston, Tex., Apr. 1, 1950.

8. Standard Handbook for Electrical Engineers (book), Archer E. Knowlton, editor. Mc-Graw-Hill Book Company, New York, N. Y., 8th ed., 1949.

9. Report of Correlating Committee on Cathodic Protection. National Association of Corrosion Engineers, Houston, Tex., July, 1951.

Discussion

F. E. Kulman (Consolidated Edison Company of New York, Inc., New York, N. Y.): Mr. Zastrow shows that the corrosion of steel anchor rods and anchors may be accelerated if they are connected electrically to other buried objects. Under certain circumstances, the corrosion of the steel anchors may be severe if the other buried objects are copper or copper-covered ground rods.

In Fig. 5, showing measurements of current in the guy wires at 108 locations, it is seen that the current flowed toward the anchor in 90 per cent of the observations, i.e., in the direction to cause corrosion of the anchor. Conversely, in Fig. 6, showing measurements of current flowing in connections between copper-covered ground rods and the neutral conductor, the current flowed from the ground rods to the neutral in all cases. It is thus evident that the use of copper-covered ground rods causes a galvanic current to circulate, resulting in the corrosion of the steel anchors. The actual rate of corrosion is influenced by the soil resistivity and the relative areas of buried copper and steel; and installations of more than the minimum number of copper ground rods may be made only at the risk of corroding other buried structures.

Mr. Zastrow's conclusion that galvanized steel ground rods should be used for replacements seems to be a sound solution. Not only does this type of rod meet the requirement of not causing corrosion on other structures but by virtue of the zinc coating the rod is self-protecting so long as the galvanizing lasts.

The author describes the difficulties of making measurements of currents in guy wires and reports no success in obtaining a correlation between anchor corrosion and the potential between neutral and the earth. In the case of underground pipe lines, the measurement problem is also difficult. The severity of corrosion has been found not to correlate well with potential difference between pipe and earth, nor with soil resistivity. There is a growing interest in the measurement of the reduction-oxidation (redox) potential of the soil as an indicator of soil corrosivity. Research[1] resulted in the preliminary design of a soil probe for measuring the redox potential, and apparently good correlation was found between the soil redox potential and the severity of pipe pitting. I would like to ask Mr. Zastrow if any attempt has been made to measure the redox potential of the soils in which the steel anchors have corroded and, if so, has any correlation been found between the redox measurement and the corrosion.

Reference

1. Anaerobic Corrosion of Iron in Soil, R. L. Starkey, K. M. Wight. American Gas Association, New York, N. Y., 1945.

F. E. Leib (Copperweld Steel Company, Glassport, Pa.): The author outlines the theoretical corrosion situation resulting from copper-zinc couples in moist earth, but his conclusion that buried copper is partly responsible for the corrosion of underground zinc-coated structures seems far from established. For example, there is no question that a potential difference exists between a copper electrode and a zinc electrode immersed in a suitable electrolyte. However, when the electrodes are electrically connected and a current flows, polarization occurs very rapidly and current diminishes to a negligible value, usually within minutes or seconds after the connection is made.

This can be demonstrated in one's own backyard. If the soil is moist, a current of perhaps 20 or 30 ma may flow at the moment of connection between two dissimilar rods, but the current will fall off to nothing within a very short time. In dry or poor conducting soils the current flow is rarely appreciable, even at the moment of connection, so that in either case the continuing current is of no practical significance in a corrosion sense.

From the simple type of test just described, one reaches the conclusion that the role of buried copper in causing zinc corrosion is negligible. This seems to be further substantiated by the author's data in which the greatest number of currents measured were very low, in the range of 0 to 2 ma. It is also significant that Mr. Zastrow recorded a number of negative readings which no amount of theorizing could associate with buried copper.

Actually, copper, zinc, and earth make a very poor battery. To maintain the potential difference between two metals, as in a dry cell, very careful attention must be paid to the electrolyte so as to minimize polarization. Even so, the life of an ordinary dry battery on short circuit is extremely limited. The tendency to associate zinc corrosion with buried copper seems further questioned in the fact that the trouble is geographically localized while the concentration of copper-type ground electrodes is in no sense geographical. In fact, there are many more copper-type grounding electrodes in the eastern states where no difficulty was reported.

Zinc-coated underground structures are more subject to corrosion than copper, as zinc is not as protective. Grounding electrodes are placed for protection and safety and must not be uncertain or unreliable. It therefore seems that any change away from copper-type grounding would be a step backward, defeating the safe grounding practices which have developed over the years, and introducing a new problem of corrosion in rustable grounding electrodes.

Further study is indicated, aimed at properly locating the source of the troublesome stray currents. Possibilities are pipe-line cathodic protection systems, surface conditions of electrodes, soil conditions in the area, irregular a-c wave shapes, unestablished rectifier actions, electric railways, static currents, etc. The least desirable course of action would be to experiment with grounding electrodes which are likely to corrode underground.

O. W. Zastrow: The contributions of Mr. Kulman and Mr. Leib are sincerely appreciated. I have not attempted any measure-

ments of the redox potentials referred to by Mr. Kulman. Any field measurement that correlates well with corrosion experience would surely be of value, and the redox potential measurement may well deserve further trial. I agree that the severity of corrosion fails to correlate well with soil resistivity or with structure potential.

The questions raised by Mr. Leib indicate that some additional details should be presented as the basis for my statement that buried copper is the most usual cause of damage to buried galvanized steel structures connected to the system neutral. Buried copper was found to be responsible for galvanic corrosion damage in detailed studies made on several electric distribution systems in each of three states: North Dakota, South Dakota, and Texas. At the time earlier studies were undertaken in Texas in 1951 it was believed (and hoped) that pipe-line protection systems or other external d-c sources might be the important source of troublesome currents. Such sources would have been much easier to deal with than a distributed source such as copper grounds. In these studies, measurements were made of direct currents in ground connections, in service neutrals (including those to rectifiers) and along the power line neutrals at various locations, and also of d-c potentials to a copper-copper sulfate reference electrode. These measurements indicated that:

1. Relatively few rectifier installations

cause appreciable flow of direct current into or along the power line neutral.

2. Where rectifiers change the d-c neutral potential, the effects usually cannot be detected as far as 1/2 mile from a rectifier in the low ground resistance areas where corrosion has been a problem.

3. So far as could be determined, most of the currents causing damage entered the system neutral at copper grounds.

Measurements made in July 1954 included a complete survey of an 8-mile section of single-phase line in Armour County, S. Dak., in an effort to account for all direct current flowing to and from the system neutral. It was found that approximately 155 ma d-c flowed toward the system neutral from 46 copper-covered ground rods. An additional 105 ma were contributed by similar grounds at nine services that did not provide any other probable d-c sources. The current was discharged into the ground at 22 anchor rods (215 ma) and at a buried galvanized pipe from an electric pump to a stock tank on one consumer's premises (45 ma).

The small magnitudes of direct currents collected by individual copper ground rods as compared with the damage resulting can be misleading. This is illustrated by the example just given, where the average current per ground rod is 3.4 ma under conditions of relatively severe corrosion. It is believed that copper grounds may cause a substantial increase in the rate of corrosion

of buried steel even in locations where galvanic corrosion is not thought a problem.

With reference to negative current readings in guys, it should be noted that galvanic effects between buried structures are not necessarily stopped when all structures are of the same kind of metal. This is pointed out in the paper. However, it is believed that the use of all galvanized steel (or all copper) underground will reduce corrosion rates so that serious problems should not arise on multigrounded electric systems.

Referring to the corrosiveness of zinc-coated steel as compared with copper-covered steel, tests by Denison and Romanoff (refs. 4 and 5 of the paper) indicate that heavily galvanized (3 ounces per square foot) steel may have better corrosion resistance than copper-covered rods with the coating thicknesses now commercially available if the copper is buried without benefit of cathodic protection from steel. The experience of Rural Electrification Administration borrowers over a period of more than 15 years indicates generally satisfactory performance of 5/8-inch galvanized steel anchor rods with a 2 ounce coating thickness, except where dissimilar metals or stray currents caused abnormally rapid corrosion. These anchor rods are installed in a grounding arrangement (described in the paper) that would tend to cause more rapid corrosion than that on an all-steel-grounded system. On the basis of experience, it appears that the presently used 2-ounce coating is adequate.

Frequency Response from Experimental Nonoscillatory Transient-Response Data

H. THAL-LARSEN
ASSOCIATE MEMBER AIEE

Synopsis: Curves are derived and presented which provide a ready means for approximating transfer functions with time constants up to three in number, and also dead time, from the experimental nonoscillatory transient response produced by a step change. The transfer function containing the time constants and dead time may be used to obtain the corresponding frequency response. These curves then are a graphical aid for converting transient response into frequency response.

FREQUENCY response from experimental transient-response data is a most sought-after transformation. Basically, the transient response of a linear system is governed by the roots of its characteristic equation. This suggests that the transient response may be used to find these roots when the equation for the system is not known. If the roots so found are real, routine

graphical procedures yield the corresponding frequency response. Complex roots are not considered here. This procedure is, of course, more often used in the reverse direction, i.e., transient response from frequency response, for two fundamental reasons: 1. the frequency response is of considerable usefulness for synthesis, and 2. methods available for resolving the transient response on paper are not always sufficiently accurate or practical. Nevertheless, from an experimental point of view, the transient-response method for finding the parameters of a linear system has appeal because of the relative ease with which the transient may be obtained. In addition, the transient is a convenient clue to the band of frequencies which must be investigated in a frequency-response analysis of the system.

Three relatively simple methods are

available for analyzing, on paper, an experimental nonoscillatory transient response. The oldest and most well-known method consists of plotting the logarithm of the error (1-transient response) against time on semilog paper. If the largest time constant is clearly dominant, the resultant curve will eventually approach a straight line with negative slope corresponding to the magnitude of the dominant time constant. Whether dead time is present or not, the straight line may be extrapolated backwards until it intersects the zero time axis at a point which may be labeled 100 per cent. A downward shift of 63 per cent, on the logarithmic scale, coupled with a horizontal shift to the sloping straight line, will also yield the time constant sought. The secondary time constant may be found by replotting the log difference, which exists at low values of time, between the sloping straight line and the transient. A

Table I. Time Ratios

Case No.	T_2	T_3	$\dfrac{t_1}{\tau_1}$	$\dfrac{t_2}{\tau_1}$	$\dfrac{t_3}{\tau_1}$	$\dfrac{t_2-t_1}{\tau_1}$	$\dfrac{t_3-t_1}{\tau_1}$	$\dfrac{t_3-t_1}{t_2-t_1}$	$\dfrac{t_2-t}{t_2-t_0}$
1	0	0	0.106	0.510	1.610	0.404	1.504	3.725	2.95
2	0.1	0	0.193	0.615	1.715	0.422	1.522	3.61	2.48
	0.2	0	0.252	0.723	1.832	0.471	1.580	3.355	2.19
	0.3	0	0.301	0.823	1.964	0.522	1.663	3.185	2.02
	0.4	0	0.343	0.915	2.103	0.572	1.760	3.08	1.92
	0.5	0	0.380	1.000	2.248	0.620	1.868	3.01	1.87
	0.6	0	0.415	1.083	2.396	0.668	1.981	2.965	1.83
3	1.0	0	0.530	1.375	2.994	0.845	2.464	2.92	1.795
4	0.1	0.1	0.280	0.720	1.820	0.440	1.540	3.50	2.14
	0.2	0.2	0.410	0.930	2.055	0.520	1.645	3.16	1.77
	0.3	0.3	0.518	1.125	2.313	0.607	1.795	2.96	1.595
	0.4	0.4	0.613	1.310	2.585	0.697	1.972	2.83	1.51
	0.5	0.5	0.704	1.484	2.863	0.780	2.159	2.77	1.455
	0.6	0.6	0.795	1.653	3.143	0.858	2.348	2.74	1.42
5	1.0	1.0	1.100	2.286	4.275	1.186	3.175	2.68	1.39
6	1	0.1	0.625	1.476	3.098	0.851	2.473	2.905	1.675
	1	0.2	0.706	1.580	3.210	0.874	2.504	2.87	1.585
	1	0.3	0.772	1.682	3.326	0.910	2.554	2.81	1.52
	1	0.4	0.831	1.777	3.450	0.946	2.619	2.77	1.475
	1	0.5	0.887	1.870	3.583	0.983	2.696	2.74	1.44
	1	0.6	0.936	1.963	3.719	1.027	2.783	2.72	1.42
7	0.6	0.1	0.498	1.186	2.500	0.688	2.002	2.93	1.69
	0.6	0.2	0.571	1.282	2.614	0.711	2.043	2.88	1.595
	0.6	0.3	0.631	1.379	2.736	0.748	2.105	2.82	1.525
	0.6	0.4	0.690	1.470	2.865	0.780	2.175	2.79	1.48
	0.6	0.5	0.745	1.562	3.000	0.817	2.255	2.76	1.445
	0.5	0.1	0.467	1.105	2.360	0.638	1.893	2.97	1.715
	0.5	0.2	0.543	1.204	2.472	0.661	1.929	2.92	1.60
	0.5	0.3	0.607	1.302	2.593	0.695	1.986	2.86	1.53
	0.5	0.4	0.662	1.395	2.724	0.733	2.062	2.81	1.48
	0.4	0.1	0.430	1.018	2.208	0.588	1.778	3.02	1.75
	0.4	0.2	0.505	1.120	2.324	0.615	1.819	2.96	1.625
	0.4	0.3	0.566	1.218	2.449	0.652	1.883	2.89	1.55
	0.3	0.1	0.390	0.927	2.068	0.537	1.678	3.12	1.81
	0.3	0.2	0.462	1.027	2.185	0.565	1.723	3.05	1.68
	0.2	0.1	0.341	0.832	1.938	0.491	1.597	3.25	1.92

Fig. 1. Normalized curves yielding time for 10-per-cent transient response corresponding to combinations of various time constants

Fig. 2. Normalized curves yielding time for 40-per-cent transient response corresponding to combinations of various time constants

repetition of these procedures will then yield the secondary time constant.

Oldenbourg and Sartorius[1] use the maximum slope and the location of the inflection point of the experimental transient-response curve to establish a time ratio which, by means of a simple curve, will yield the time constants of the equivalent second-order system. St. Clair[2] has suggested a method which consists of plotting the experimental transient response, with the exception of the first 1 per cent, on log-log paper. The shape of the resultant curve is then matched against the shape of calculated curves on a template. Corresponding to these calculated transient-response curves on log-log paper are the calculated frequency-response curves, also plotted on templates. Thus, with a sufficiently wide selection of templates, the conversion from transient to frequency response is readily effected.

This paper adds one more method to those already available for resolving an experimental nonoscillatory transient response produced by a step change. In application it is quick and convenient.

Derivation of Curves

The transfer function for a noninteracting 3-time-constant system may be written, assuming a final steady-state response equal to unity

$$G(S) = \frac{1}{(\tau_1 S+1)(\tau_2 S+1)(\tau_3 S+1)} \quad (1)$$

where τ_1, τ_2, and τ_3 are time constants arranged in decreasing order of magnitude. Dividing the time constants by τ_1, the largest time constant, will normalize equation 1. Assuming again a final steady-state response equal to unity, the normalized transfer function may be written

$$G_n(S) = \frac{1}{(S+1)(T_2 S+1)(T_3 S+1)} \quad (2)$$

where $T_2 = \tau_2/\tau_1$ and $T_3 = \tau_3/\tau_1$ represent relative or dimensionless time constants.

Seven cases may be considered for the transient response of the normalized system resulting from the application of a unit step change:

1. $T_2 = T_3 = 0$
2. $0 < T_2 < 1$, $T_3 = 0$
3. $T_2 = 1$, $T_3 = 0$
4. $T_2 = T_3$, $0 < T_3 < 1$
5. $T_2 = T_3 = 1$
6. $T_2 = 1$, $0 < T_3 < 1$
7. $0 < T_2 < 1$, $0 < T_3 < 1$, $T_2 > T_3$

The corresponding transient responses are given by

$$c_1(\theta) = 1 - \epsilon^{-\theta} \quad (3)$$

$$c_2(\theta) = 1 - \frac{\epsilon^{-\theta}}{1-T_2} + \frac{T_2\epsilon^{-\theta/T_2}}{1-T_2} \quad (4)$$

$$c_3(\theta) = 1 - (1+\theta)\epsilon^{-\theta} \quad (5)$$

$$c_4(\theta) = 1 + \frac{\theta\epsilon^{-\theta/T_2}}{1-T_2} + \frac{T_2(2-T_2)\epsilon^{-\theta/T_2}}{(1-T_2)^2} - \frac{\epsilon^{-\theta}}{(1-T_2)^2} \quad (6)$$

$$c_5(\theta) = 1 - \left(1+\theta+\frac{\theta^2}{2}\right)\epsilon^{-\theta} \quad (7)$$

$$c_6(\theta) = 1 - \frac{\theta\epsilon^{-\theta}}{1-T_3} - \frac{(1-2T_3)\epsilon^{-\theta}}{(1-T_3)^2} - \frac{T_3^2\epsilon^{-\theta/T_3}}{(1-T_3)^2} \quad (8)$$

$$c_7(\theta) = 1 - \frac{\epsilon^{-\theta}}{(1-T_2)(1-T_3)} + \frac{T_2^2\epsilon^{-\theta/T_2}}{(1-T_2)(T_2-T_3)} - \frac{T_3^2\epsilon^{-\theta/T_3}}{(1-T_3)(T_2-T_3)} \quad (9)$$

Fig. 3. Normalized curves yielding time for 80-per-cent transient response corresponding to combinations of various time constants
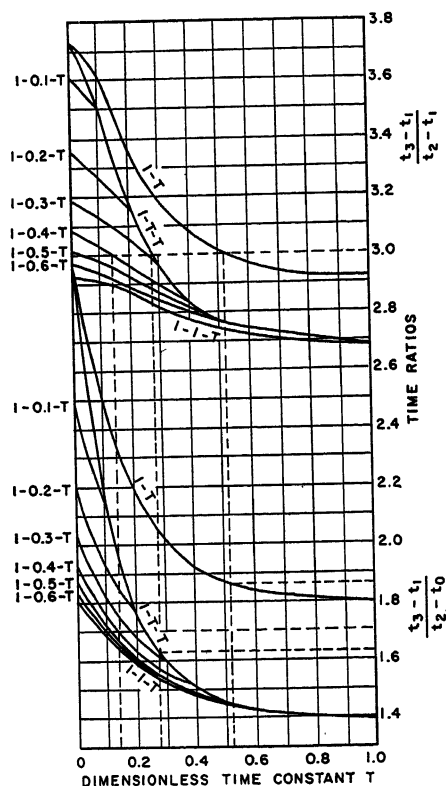
Fig. 4. Normalized curves yielding time-interval ratios of the transient response corresponding to combinations of various time constants

where θ = dimensionless time and is equal to actual time t divided by τ_1.

The problem consists of setting each of equations 3 through 9 in turn equal to 0.1, 0.4, and 0.8 respectively, representing 10-, 40-, and 80-per-cent response, and solving by trial and error for the corresponding values of θ, labeled t_1/τ_1, t_2/τ_1, and t_3/τ_1 respectively, for various values of T_2 and T_3. The solution is tedious and time-consuming. Final results for t_1/τ_1, t_2/τ_1, and t_3/τ_1 are listed in Table I and have been plotted in Figs. 1, 2, and 3. The subsequently computed quantities $(t_3-t_1)/\tau_1$, $(t_3-t_1)/(t_2-t_1)$, and $(t_3-t_1)/(t_2-t_0)$ are also listed in Table I and have been plotted in Figs. 4 and 5. The quantity t_0 is used to represent dead time for the experimental transient response. Since this term is zero for the foregoing computed transient responses, it follows that $(t_3-t_1)/(t_2-t_0)=(t_3-t_1)/t_2$ for these computed quantities.

Method of Using Curves

In applying the curves, the method consists of taking measurements on three primary points of the experimental non-oscillatory transient produced by a step change. These have been arbitrarily located at the 10-, 40-, and 80-per-cent response levels. The time corresponding to these points is noted and designated t_1, t_2, and t_3 respectively. The dimensionless ratio $(t_3-t_1)/(t_2-t_1)$ and the time (t_3-t_1), used in conjunction with Figs. 4 and 5, are sufficient to select roots which will reproduce a transient through the primary points just given. It should be noted that dead time does not affect the results. If dead time is present and estimated to be equal to time t_0, the dimensionless ratio $(t_3-t_1)/(t_2-t_0)$ enables the selection of those particular roots which will also reproduce the first 10 per cent or toe of the transient. For convenience, the curves in this paper have been plotted in terms of time constants, i.e., the reciprocals of the roots, with the sign positive.

Example

An example is given to illustrate the use of the curves. Assume an actual transient response with $t_1=0.97$ minutes (min.), $t_2=2.14$ min., $t_3=4.47$ min., as shown in Fig. 6, and with, perhaps, a certain amount of dead time t_0. Such a transient might be encountered in the chemical industry. Then

$$\frac{t_3-t_1}{t_2-t_1}=\frac{3.50}{1.17}=3.00$$
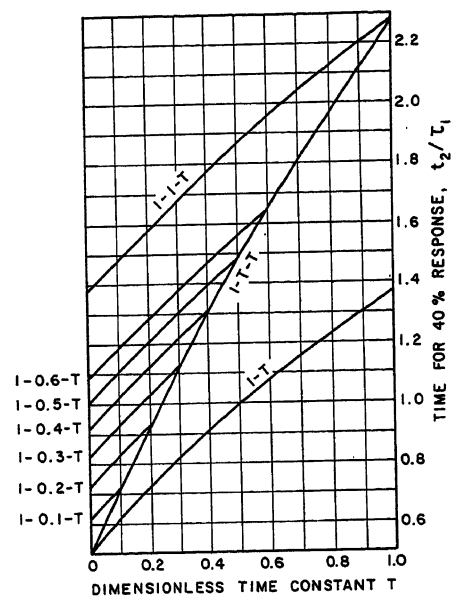
Fig. 5. Normalized curves yielding the time interval between 10- and 80-per-cent response on the transient corresponding to combinations of various time constants

Step 1. Enter Fig. 4 at 3.00 and note that this line crosses several curves, allowing the choice of various combinations of the dimensionless or relative time constant T. Choosing three of these combinations:

From curve 1-T-T: 1, 0.275, 0.275.
From curve 1-0.4-T: 1, 0.4, 0.135.
From curve 1-T: 1, 0.520.

Step 2. Enter Fig. 5 with the dimensionless time constants of step 1 and determine the associated dimensionless time $(t_3-t_1)/\tau_1$:

From curve 1-T-T for $T=0.275$: 1.755.
From curve 1-0.4-T for $T=0.135$: 1.790.
From curve 1-T for $T=0.520$: 1.890.

Step 3. The time (t_3-t_1) from the actual transient divided by the dimensionless time $(t_3-t_1)/\tau_1$ of step 2 yields the conversion factor by which the relative time constants, found in step 2, must be multiplied to obtain the actual time constants designated by τ. For the assumed actual transient $(t_3-t_1)=3.50$ min. Then:

$(3.50/1.755)[1, 0.275, 0.275]$:

$\tau_1=1.995$ min.
$\tau_2=\tau_3=0.549$ min.

$(3.50/1.790)[1, 0.4, 0.135]$:

$\tau_1=1.955$ min.
$\tau_2=0.782$ min.
$\tau_3=0.264$ min.

$(3.50/1.890)[1, 0.520]$:

$\tau_1=1.850$ min.
$\tau_2=0.963$ min.
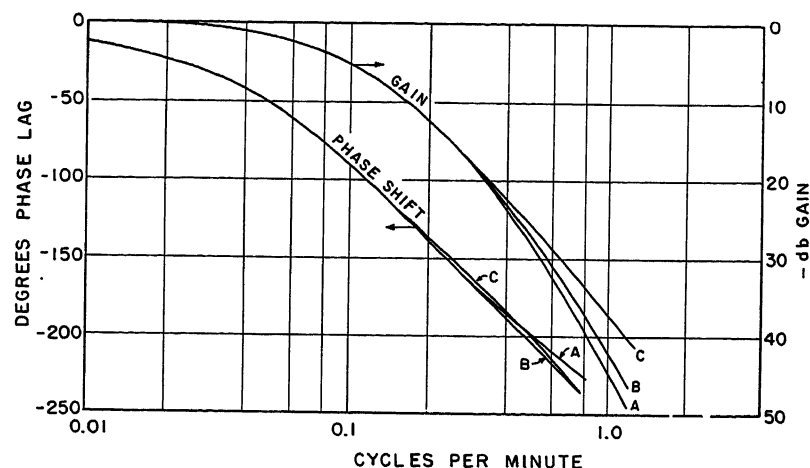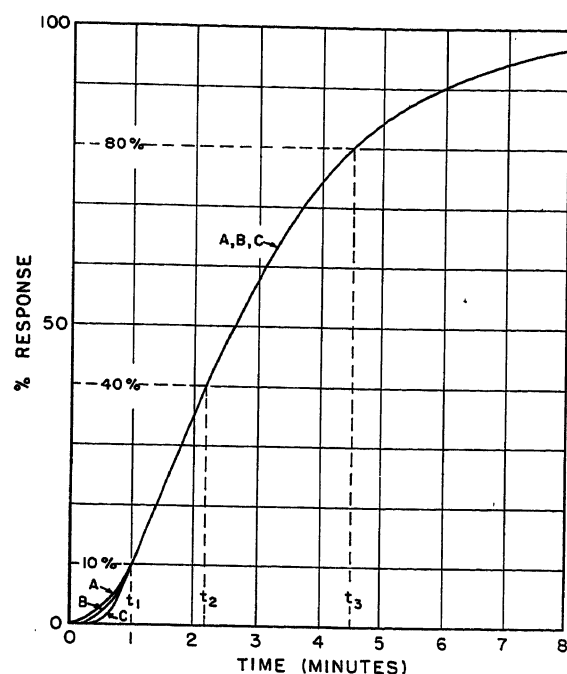
Fig. 6 (left). Transient-response curves corresponding to frequency-response curves of Fig. 7

Fig. 7 (above). Frequency-response curves corresponding to transient-response curves of Fig. 6

Step 4. Also enter Fig. 2 with the dimensionless time constants from step 2 and determine the associated dimensionless time t_2/τ_1:

From curve 1-T-T for $T=0.275$: 1.075.
From curve 1-0.4-T for $T=0.135$: 1.055.
From curve 1-T for $T=0.520$: 1.018.

Step 5. Application of the conversion factor from step 3 to the quantities found in step 4 yields the actual time t_2 if no dead time t_0 is present. Thus:

(3.50/1.755)(1.075)=2.142 min.
(3.50/1.790)(1.055)=2.060 min.
(3.50/1.890)(1.018)=1.885 min.

Step 6. Dead time t_0 can now be found by subtracting the time given in step 5 from the actual value for t_2 which, in this case, is 2.14 min. Therefore:

$t_0=2.14-2.142=0$
$t_0=2.14-2.060=0.080$ min.
$t_0=2.14-1.885=0.255$ min.

Step 7. The following three transfer functions will therefore reproduce the actual transient, at least over the interval from 10- to 80-per-cent response:

$$G(S) = \frac{1}{(1.995S+1)(0.549S+1)^2} \quad \text{(A)}$$

$$G(S) = \frac{\epsilon^{-0.080S}}{(1.955S+1)(0.782S+1)(0.264S+1)} \quad \text{(B)}$$

$$G(S) = \frac{\epsilon^{-0.255S}}{(1.850S+1)(0.963S+1)} \quad \text{(C)}$$

The transients corresponding to these transfer functions are shown in Fig. 6, and the corresponding frequency response in Fig. 7. (It is noted as a matter of interest that the summations of the time constants, plus dead time, if present, are substantially equal for these transfer functions.)

Step 8. Suppose that an examination of the start of the transient suggests an estimated value of dead time $t_0=0.08$ min. Then

$$\frac{t_3-t_1}{t_2-t_0}=\frac{3.50}{2.06}=1.70$$

Step 9. Enter Fig. 4 at 1.70 and note that curve 1-0.4-T is the only one which satisfies simultaneously this value and the previously determined value, $(t_3-t_1)/(t_2-t_1)=3.00$, for the same T. The second transfer function of step 7 will therefore reproduce the actual transient the best.

Note that a time ratio $(t_3-t_1)/(t_2-t_1)$ smaller than 2.92 indicates that the transient response cannot be reproduced by a 2-time-constant system; nor by a 3-time-constant system if the ratio is less than 2.68. Note also that only a limited range of choices exists for the ratio $(t_3-t_1)/(t_2-t_0)$ once the ratio $(t_3-t_1)/(t_2-t_1)$ has been established. Thus the range for dead time is also limited.

If the actual transient has such a shape that the ratio $(t_3-t_1)/(t_2-t_1)$ falls outside the region covered by Fig. 4, it can be concluded that the system cannot be represented by a 3-time-constant system. However, since two or three time constants may provide a sufficiently accurate approximation under certain circumstances, it is suggested that an exponential curve be "faired" through three primary check points

which will produce a satisfactory time ratio. It may also be necessary to fair the substitute curve into an assumed 100-per-cent response level if the actual transient is unsatisfactory in that respect.

Transients of systems having more than three time constants can still be approximated by a transfer function having only three time constants, provided that the time constants of the higher order system are not all nearly of the same magnitude. A rough guide, which has not been tested, is probably the rule that the summation of the time constants of the higher order system should be less than three times the largest time constant.

Accuracy

An examination of Fig. 7, which is the desired end product, namely the frequency response, shows that the amplitude responses for all three transfer functions are so close to each other as to be practically indistinguishable until the amplitude ratio has dropped to 10 per cent at -20 decibels. At 0.82 cycles per min. the amplitude ratio for curve A has dropped to 1 per cent and for curve C, representing the greatest deviation, to 2 per cent, a difference which is hardly significant. Also it will be noted that the phase-angle characteristics lie along each other except at the extreme high-frequency ends of the curves. Thus one need not be greatly concerned if the transfer functions deviate from each other slightly.

The large number of apparently significant figures in the section entitled "Example" have been used for computational purposes. They should not be

construed as an index to the accuracy of this method. If curves are replotted on millimeter paper from the figures in Table I, more accurate results can be obtained than are possible with the curves published in this paper.

Accuracy is, of course, dependent upon the quality of both the step change and the transient obtained. In other words, if the step is not sharp and of infinite slope, or if the recording system is inaccurate, or if the transient is very much affected by nonlinearities in the system or by extraneous disturbances, poor results ensue. The accuracy of the results is also affected by the relative magnitude of the roots because of the decreasing slope of the curves in Fig. 4 as the time constants approach unity.

Conclusions

1. The method presented here affords a means of approximating the transfer func-tions of a physical system (either linear or nonlinear), from simple measurements on its nonoscillatory transient response, by a 3-time-constant-plus-dead-time type of transfer function. The latter may then be used to obtain the frequency response of the system.

2. Estimates of dead time, if present in the experimental transient response, are limited in range by virtue of the curves of Fig. 4. An inaccurate estimate of dead time within this limited range does not materially affect the conversion from transient response to equivalent frequency response.

3. By their conformity the curves of Figs. 6 and 7 suggest that the first 10 per cent or toe of the transient-response curve may not be as important as previously thought for many applications.

4. Slope and spacing of the curves of Figs. 1, 2, and 3 show that the second and third largest time constants influence the course of the transient approximately in direct proportion to the percentage level of the response. Thus measurements made upon the experimental transient between the 10- and 80-per-cent response levels are more conducive to an accurate determination of an equivalent transfer function than measurement made upon the first 10 per cent or toe of the transient.

5. It may be possible to extend the type of analysis developed in this paper to systems of higher order. Complex roots, and hence oscillatory transients, may yield to a similar treatment.

6. This method may be useful for very slow processes, which make a frequency-response analysis extremely time-consuming, or for ultrafast components, such as certain pressure gauges, which make a frequency-response analysis impractical because of the speed and pressure limitations introduced by a pneumatic sine-wave generator.

References

1. THE DYNAMICS OF AUTOMATIC CONTROLS (book), R. C. Oldenbourg, H. Sartorius. American Society of Mechanical Engineers, New York, N. Y., 1948, pp. 77–78.

2. STEP RESPONSE AS A SHORT CUT TO FREQUENCY RESPONSE, D. W. St. Clair. *Proceedings*, Instrument Society of America, Pittsburgh, Pa., vol. 7, 1952, pp. 96–101.

Discussion

Yasundo Takahashi (Massachusetts Institute of Technology, Cambridge, Mass.):
The author is to be congratulated in completing such a laborious work. In addition to the error in the high-frequency range, mentioned by the author in the section on "Accuracy," there may be a possible error in the medium-frequency band if there are any derivative time constants in the system. Consider, e.g., the system whose transfer function is

$$G(S) = \frac{1+\tau_4 S}{(1+S)(1+\tau_2 S)(1+\tau_3 S)}$$

and whose time constants fall within the range of Fig. 4. If this derivative function is replaced by a system given by equation 1 of the paper, it is conceivable that the error might be considerable. However, the error was extremely small in my trials; thus one may conclude that the author's choice of the 80-per-cent response is most proper.

The author's third conclusion might be in error for systems where there is a single integration plus a dead time (case 1, Fig. 8) or systems with two integrations (case 2, Fig. 8). However, this is a hypothetical situation and outside the scope of the paper.

Thomas J. Higgins (University of Wisconsin, Madison, Wis.):
This paper is well written. The purpose is stated clearly; some account is given of the simpler methods already in use, so that the interested reader can ascertain in what manner the author's approach differs from these; the development of the basic theory of his method is clearly delineated; explanation of use of the curves stemming from this development is well stated; the numerical problem excellently illustrates the typical course of solution; the discussion of accuracy is helpful; and the essential values of the paper are well summarized in the conclusions.

R. M. Saunders (University of California, Berkeley, Calif.):
While in many cases frequency response is the most sought-after result, another aspect of the author's paper should not be overlooked: that of ascertaining a transfer function of a block by experimental means. The author shows that from the transient response of any nonoscillatory system a new system can be synthesized in transfer function form. This is most significant in terms of the working out by analytical means of compensating schemes, establishing conditions for stability, setting gains, ascertaining sensitivities, etc., for such systems.

H. Thal-Larsen:
I appreciate the comments and the further elucidation of my paper by the discussers.

Professor Takahashi suggests that an unsuspected derivative time constant, say, τ_4, may introduce an error into the medium-frequency band. The effect of τ_4 is to raise the transient-response curve, the maximum vertical shift occurring at the inflection point of the transient obtained with $\tau_4 = 0$. If τ_4 is sufficiently large compared to τ_1 the resultant transient will rise momentarily above the 100-per-cent response level. This distortion should serve to warn of the existence of a derivative time constant. Resolution of the transient is then outside the scope of this paper. If τ_4 is sufficiently small compared to τ_3, only the high-frequency band is affected. This region is of relatively little importance.

Our concern over the effect of τ_4 is then reduced to the medium-frequency band, i.e., we must examine the effect of values of τ_4 in the region covered by τ_1, τ_2, and τ_3, as pointed out by Professor Takahashi. Values of τ_4 close to either τ_1, τ_2, or τ_3 will result in a transient which may be approximated by a 2-time-constant delay and which therefore can be resolved as shown in the paper. The error in the medium-frequency band is then negligible.

Values of τ_4 within the region of concern but not close to either τ_1, τ_2, or τ_3 will result in a transient which may or may not be sufficiently distorted to prevent resolution. Suppose that the actual but unknown system transfer function

$$\frac{(\tau_4 S+1)e^{-t_0 S}}{(\tau_1 S+1)(\tau_2 S+1)(\tau_3 S+1)}$$

can be resolved, by means of the transient, into the transfer function

$$\frac{e^{-t_0 S}}{(\tau_5 S+1)(\tau_6 S+1)(\tau_7 S+1)}$$

The transient is not altered by modifying the transfer function to

$$\frac{(\tau_4 S+1)e^{-t_0 S}}{(\tau_4 S+1)(\tau_5 S+1)(\tau_6 S+1)(\tau_7 S+1)}$$
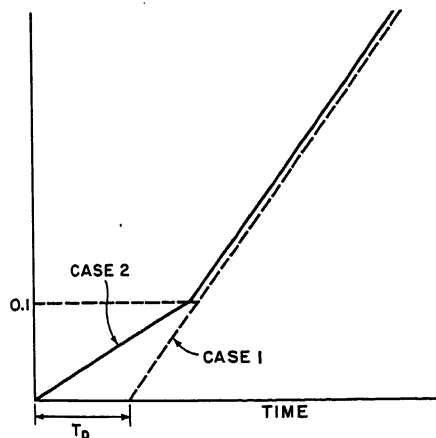
Fig. 8 Systems showing simple integration plus a dead time and two integrations

Then τ_4 must be of such a magnitude that it can be used as one of the time constants of a 4-time-constant delay approximately equivalent to the actual 3-time-constant delay. In this case also the error produced by τ_4 in the medium-frequency band will be negligible. Extension of the scope of this paper to 4-time-constant systems would allow the foregoing analysis to be tested on a quantitative basis.

The presence of integration terms in the actual system transfer function would, of course, remove the transient from the scope of this paper. The unstabilizing influence of these terms should, however, be sufficiently apparent in the transient with the result that their presence should not remain unnoticed.

A Method for Evaluating Nonlinear Servomechanisms

M. V. MATHEWS
ASSOCIATE MEMBER AIEE

Synopsis: A method is presented for determining a valid mathematical description of a class of nonlinear servomechanisms from measurements of the servo response to a set of input signals. In the sense used here, a valid mathematical description means a description that predicts the response of the servomechanism to a given class of inputs with a desired degree of accuracy. The set of input signals may consist of either sinusoids or Gaussian random signals. The response of the servomechanism is measured as a function of the amplitude and frequency composition of the input. The types of servomechanisms treated are conventional, single-loop feedback systems having a zero energy-storage nonlinear element, which may be located anywhere in the forward part of the loop. The evaluation procedure determines the transfer functions of the linear elements on both sides of the nonlinear element and the response function of the nonlinear element. The characteristics of the nonlinear and linear element must be such that only a small amount of nonlinear distortion exists at the input to the nonlinear element.

THE servo-evaluation problem considered in this paper is the problem of determining the characteristics of a servo by measuring its response to a test input signal. For example, sinusoidal, transient, or random-signal input-response data may be used to find the transfer function of a linear system.

Evaluation is one of the basic problems in the servo field, and the need for better procedures, particularly for nonlinear servos, has long been recognized.[1] Evaluation methods that are based on purely linear theory, such as the measurement of transient or sinusoidal performance, have significant limitations. These limitations arise from nonlinear effects, which are inherent in all practical servos. In the most ideal cases, careful control of the amplitude of the test signals is necessary to ensure linear operation of the servo under test. In many practical

cases, the servo exhibits no linear operating region owing to the overlapping effects of low-level nonlinearities, such as dead space, and high-level nonlinearities, such as saturation. Interpretation of data by a linear-evaluation method to obtain a description of the linear parts of servo is almost impossible in this instance.

Two methods are presented which avoid the main limitations of the linear evaluation techniques by taking into account the inherent nonlinear performance of the servo. The first method involves the use of sinusoidal data for evaluation; the second method is similar but uses random-signal data.

Evaluation with Sinusoidal Data

The sinusoidal procedure is developed from the methods for calculating analytically the response of certain types of nonlinear servos presented by Kochenburger[2] and Johnson[3] in which the nonlinear servo is approximated by a quasi-linear system. The term quasi-linear as used in this paper refers to a linear element or system, the parameters of which are functions of the input signal to the element or system.

The servo to be evaluated must be representable by the form of the system shown in Fig. 1(A) which is that of a conventional single-loop servo. The forward part of the loop is made up of two linear elements with transfer functions $A(s)$ and $B(s)$ and a nonlinear element with an instantaneous response function $f(e_1')$. The output of the nonlinear element e_2' is given by the function

$$e_2' = f(e_1') \tag{1}$$

where e_1' is the input to the nonlinear element.

The first step in the evaluation procedure is to determine the best quasi-linear approximation, Fig. 1(B), to the

nonlinear servo by replacing the nonlinear element in the servo by its quasi-linear approximation. The quasi-linear representation of the $f(e_1')$ nonlinear element is a simple gain function[2,3] $F(E_1)$ which is determined from $f(e_1')$ by the equation

$$F(E_1) = \frac{2}{\pi E_1} \int_{-\pi/2}^{+\pi/2} f(E_1 \sin \theta) \sin \theta \, d\theta \tag{2}$$

Thus, in the quasi-linear system

$$\frac{e_2}{e_1} = F(E_1) \tag{3}$$

where e_1 and e_2 are the input and output of the quasi-linear representation, and e_1 is a sinusoid with peak amplitude E_1.

The assumption is made that the quasi-linear system is a sufficiently good approximation to the nonlinear servo so that data from the nonlinear servo may be used to evaluate the elements in the quasi-linear system. The transfer functions $A(s)$ and $B(s)$ and the quasi-linear function $F(E_1)$ are evaluated from input-response data taken experimentally from the servo. The evaluation of the servo is completed by noting that $A(s)$ and $B(s)$ are also the transfer functions of the linear servo elements and by determining the instantaneous response function $f(e_1')$ from the quasi-linear function $F(E_1)$. This latter determination is obtained by the solution of equation 2, an integral equation, the solution of which is discussed in Appendix I.

The principal limitation to the evaluation procedure is the required assumption that a good quasi-linear approximation to the nonlinear servo may be found. Thus, only systems which can be effectively analyzed by the Kochenburger-Johnson techniques can be effectively evaluated. These systems must generate only a small amount of nonlinear distortion. However, the point should be made that these systems can have elements that introduce extremely non-

Fig. 1. Form of servo to be evaluated. The notation p denotes the differential operator

A—Nonlinear servo
B—Quasi-linear equivalent
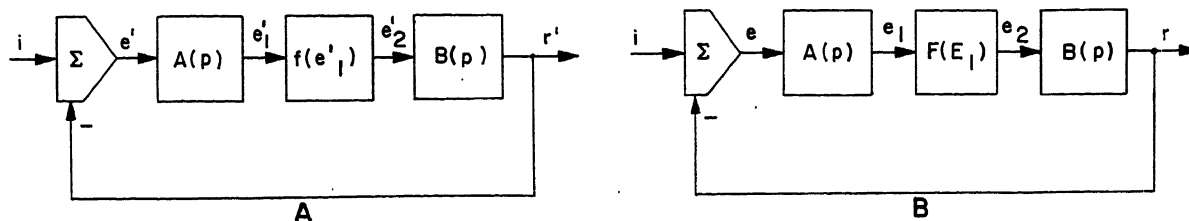
linear functions, e.g., sharp saturation.

The experimental data are obtained by exciting the servo with a sinusoid and measuring the response as a function of the peak amplitude and frequency of the input. The instantaneous time function for a sinusoidal input applied to the servo shown in Fig. 1(A) may be written

$$i = I \cos (\omega t) \qquad (4)$$

where I is the peak amplitude of the input and ω is the angular frequency. With this input the signals in the servo may be expressed as

$$e' = E \cos (\omega t + \epsilon) + \text{nonlinear distortion in } e'$$
$$e_1' = E_1 \cos (\omega t + \epsilon_1) + \text{nonlinear distortion in } e_1'$$
$$e_2' = E_2 \cos (\omega t + \epsilon_2) + \text{nonlinear distortion in } e_2'$$
$$r' = R \cos (\omega t + \rho) + \text{nonlinear distortion in } r' \qquad (5)$$

in which the lower-case letters represent instantaneous time functions, the capital letters represent the peak amplitudes of the fundamental components of these time functions, and the Greek letters represent the phase angles of the fundamental components. All the other signal components produced by the non-linear element are lumped together as nonlinear-distortion terms.

If the input is applied to the quasi-linear approximation and the basic assumption is made that the fundamental components of the signals in the quasi-linear approximation are equal to the fundamental components of the signals in the servo, the signals in the quasi-linear approximation may be written

$$e = E \cos (\omega t + \epsilon)$$
$$e_1 = E_1 \cos (\omega t + \epsilon_1)$$
$$e_2 = E_2 \cos (\omega t + \epsilon_2) \qquad (6)$$
$$r = R \cos (\omega t + \rho)$$

The quantities in equation 6 are defined in the same manner as the corresponding quantities in equation 5.

The data for the servo evaluation consist of the ratio of the peak amplitude of the response to the peak amplitude of the error R/E, measured as a function of the amplitude of the error E and the

frequency of the input ω. A set of data from a typical servo which will be used as an example is shown in Fig. 2. These data were obtained from a simulation of the servo on an electronic differential analyzer. The forward loop of the simulated servo consists of a compensation network with transfer function $A(s)$, a saturation element with response function $f(e_1')$ which represents torque saturation in the prime mover, and a linear element with transfer function $B(s)$ which represents the linear characteristics of the prime mover.

With some servos, a direct measurement of R/E and E is feasible. With other servos, a measurement of either

the response or the error in the servo as a function of the amplitude and frequency of the input is convenient; the ratio R/E can be determined from the measured data by any of a number of well-known mapping functions, such as the log-modulus plot.[4,5]

The performance of the servo as a function of E and ω may be described by two equations

$$E_1 = |A(j\omega)| E \qquad (7)$$

$$\frac{R}{E} = |A(j\omega)B(j\omega)| F[|A(j\omega)| E] \qquad (8)$$

Equation 7 gives the input to the quasi-linear element E_1 as a function of $|A(j\omega)|$ and E. Equation 8 gives the
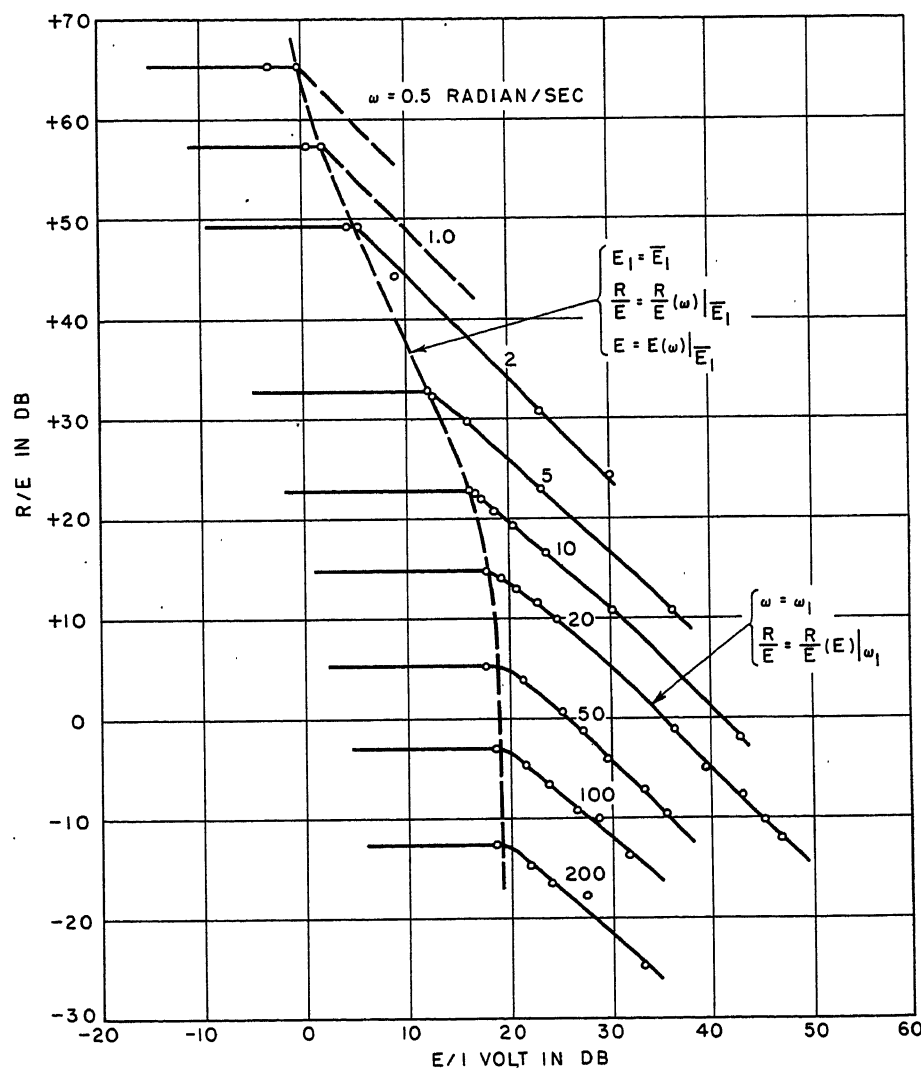
Fig. 2. Sinusoidal data

Table I. Evaluation of F(E₁)

E/1 Volt, Db	$\frac{R}{E}(E)\Big|_{\infty}$, Db	F(E₁), Db	E₁/1 Volt, Db
10	14.7	0	12
15	14.7	0	17
18	14.7	0	20
20	13.5	− 1.2	22
25	9.6	− 5.1	27
35	0.1	−14.6	37
45	−10.0	−24.7	47

ratio of response to error R/E as a function of $|A(j\omega)B(j\omega)|$ and $F(E_1)$.

One of the R/E curves shown in Fig. 2 may be used with equations 7 and 8 to evaluate the quasi-linear gain function $F(E_1)$. The frequency ω is constant along the curve, and the value of R/E may be denoted thus

$$\frac{R}{E}(E)\Big|_{\omega_1}$$

Substitution of this value into equation 8 and a solution for $F(E_1)$ yield the relation

$$F(E_1) = \frac{\frac{R}{E}(E)\Big|_{\omega_1}}{|A(j\omega_1)B(j\omega_1)|} \tag{9}$$

The value of E_1 is given by the relation

$$E_1 = |A(j\omega_1)|E \tag{10}$$

Thus, the function $F(E_1)$ is evaluated by equations 9 and 10. The two arbitrary constants $|A(j\omega_1)|$ and $|B(j\omega_1)|$ which appear in equations 9 and 10 have no effect on the input-response performance of the servo and only serve to determine the scale factors of internal signals in the forward loop. Consequently, they cannot be evaluated from input-response data, but must either be specified arbitrarily or determined from additional measurements made on the internal servo signals.

The numerical determination of $F(E_1)$ is readily carried out in the tabular form shown in Table I. Values of E and $R/E(E)|_{\omega_1}$ from Fig. 2 are tabulated

in the first two columns. Values of $F(E_1)$ and E_1 computed with equations 9 and 10 are tabulated in the last two columns. The arbitrary constants are chosen to make $F(E_1)=0$ decibels (db) for small values of E_1 and to make $F(E_1)$ begin to decrease at $E_1=20$ db. The required values are: $|A(j\omega_1)|=2$ db., $|B(j\omega_1)|=12.7$ db. The good agreement between the plot of $F(E_1)$ shown in Fig. 3 and the plot of the quasi-linear gain function of the simulated nonlinear element establishes the accuracy of the evaluation procedure. The gain function $F(E_1)$ corresponds to the sharply saturating nonlinear response function shown in Fig. 4. The determination of $f(e_1')$ from $F(E_1)$ is discussed in Appendix I.

The linear transfer functions $A(j\omega)$ and $B(j\omega)$ may be evaluated from one R/E curve along which E_1 is constant. This curve cannot be determined directly from experimental data but must be located from the curves of constant ω. The method for its location will be described subsequently. For the present, the assumption is made that one such curve has been located, as shown by the dashed line in Fig. 2. The co-ordinates of this curve may be expressed parametrically as functions of ω: Thus, the value of R/E on this curve may be defined by the relation

$$\frac{R}{E} = \frac{R}{E}(\omega)\Big|_{\bar{E}_1} \tag{11}$$

and the value of E on this curve may be defined by the relation

$$E = E(\omega)\big|_{\bar{E}_1} \tag{12}$$

Substitution of the expression for E from equation 12 into equation 7 and solution for $|A(j\omega)|$ yield

$$|A(j\omega)| = \frac{\bar{E}_1}{E(\omega)\big|_{\bar{E}_1}} \tag{13}$$

The value of $\bar{E}_1$ may be written

$$\bar{E}_1 = |A(j\omega_1)||E(\omega_1)|_{\bar{E}_1} \tag{14}$$

By combining equations 13 and 14 to eliminate $\bar{E}_1$, the relation

$$|A(j\omega)| = |A(j\omega_1)|\frac{E(\omega_1)\big|_{\bar{E}_1}}{E(\omega)\big|_{\bar{E}_1}} \tag{15}$$

is obtained to complete the evaluation of $|A(j\omega)|$.

Substituting the data from equation 11 into equation 8, $|A(j\omega)B(j\omega)|$ may be written

$$|A(j\omega)B(j\omega)| = \frac{\frac{R}{E}(\omega)\Big|_{\bar{E}_1}}{F(\bar{E}_1)} \tag{16}$$

and $F(\bar{E}_1)$ may be written

$$F(\bar{E}_1) = \frac{\frac{R}{E}(\omega_1)\Big|_{\bar{E}_1}}{|A(j\omega_1)B(j\omega_1)|} \tag{17}$$

By combining equations 16 and 17 to eliminate $F(\bar{E}_1)$, the relation

$$|A(j\omega)B(j\omega)| = |A(j\omega_1)B(j\omega_1)|\frac{\frac{R}{E}(\omega)\Big|_{\bar{E}_1}}{\frac{R}{E}(\omega_1)\Big|_{\bar{E}_1}} \tag{18}$$

is obtained for the evaluation of $|A(j\omega)B(j\omega)|$. The magnitude $|B(j\omega)|$ is obtained from the identity

$$|B(j\omega)| \equiv \frac{|A(j\omega)B(j\omega)|}{|A(j\omega)|} \tag{19}$$

and the values of $|A(j\omega)|$ and $|A(j\omega)B(j\omega)|$ from equations 15 and 18.

The computation of $|A(j\omega)|$ and $|B(j\omega)|$ is easily carried out in the tabular form illustrated in Table II. The R/E and E data from Fig. 2 are tabulated as a function of ω in the first three columns. The values of $|A(j\omega)|$, $|A(j\omega)B(j\omega)|$, and $|B(j\omega)|$ computed from equations 15, 18, and 19 are tabulated in the last three columns. The evaluated magnitudes $|A(j\omega)|$ and $|B(j\omega)|$ are plotted in Fig. 5. These magnitudes agree closely with magni-
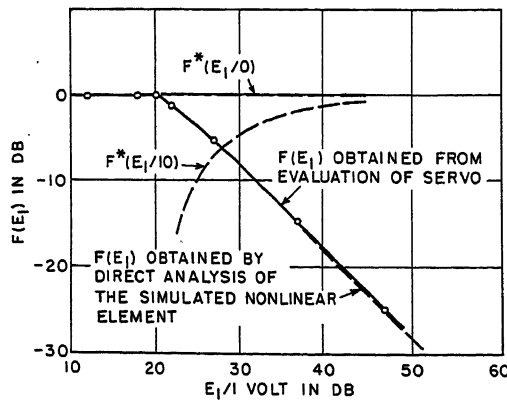
Fig. 3 (left). Quasi-linear gain function
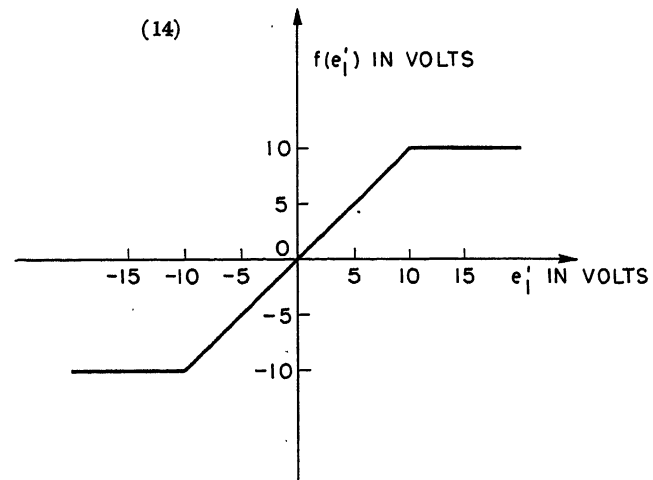
Fig. 4 (right). Nonlinear response function

Table II. Evaluation of |A(jω)| and |B(jω)|

ω, Radians per Second	$\frac{R}{E}(\omega)\big|_{\bar{E}_1}$, Db	$E(\omega)\big|_{\bar{E}_1}$/1 Volt, Db	$|A(j\omega)|$, Db	$|A(j\omega)B(j\omega)|$, Db	$|B(j\omega)|$, Db
0.5	65.3	−0.3	20.3	65.3	45.0
1	57.0	2.0	18.0	57.0	39.0
2	49.2	5.3	14.7	49.2	34.5
5	32.8	12.2	7.8	32.8	25.0
10	22.9	16.5	3.5	22.9	19.4
20	14.9	17.9	2.1	14.9	12.8
50	5.3	18.9	1.1	5.3	4.2
100	−3.1	19.0	1.0	−3.1	−4.1
200	−12.6	19.2	0.8	−12.6	−13.4

tudes of the simulated transfer functions also shown in Fig. 5, thus illustrating the accuracy of the evaluation procedure.

Equations 15 and 19 constitute a method of evaluating $|A(j\omega)|$ and $|B(j\omega)|$. In order to obtain the transfer functions $A(s)$ and $B(s)$ from the magnitude functions, some procedure such as Bode's[6] may be used to associate phase angles with the magnitude functions. In associating these phase angles, assumptions with regard to the phase characteristics of $A(s)$ and $B(s)$ must be made. These assumptions can be checked, since the total phase shift in the forward loop of the servo can be determined from the experimental data.

The location of a curve of constant E_1 is accomplished by plotting $\ln R/E$ as a function of $\ln E$ for several values of ω. Such a logarithmic plot is used in Fig. 2. If logarithms of both sides of equation 8 are taken and the substitution

$$E = \epsilon \ln E \qquad (20)$$

is made, the following relation is obtained

$$\ln \frac{R}{E} = \ln |A(j\omega)B(j\omega)| + \ln F(e^{n\,E + \ln|A(j\omega)|}) \qquad (21)$$

Because of the functional form of equation 21, the curves of $\ln R/E$ versus $\ln E$ for various values of ω must have the same shape, and any curve may be brought into coincidence with any other curve by shifting it horizontally and vertically. This similarity is illustrated in Fig. 2.

If similar points on the curves of constant ω are connected, the resulting curve is one of constant E_1. In other words, a line along which one of the curves of constant ω must be shifted in order to bring it successively into coincidence with all the other curves of constant ω

is one of constant E_1. The line $E_1 = \bar{E}_1$ is illustrated in Fig. 2.

The formation of a family of similar curves is a test of adequacy of the evaluation procedure. If the servo data form such a family of curves, the nonlinear servo can be represented by the type of nonlinear system shown in Fig. 1(A), and the servo may be approximated by the quasi-linear system of Fig. 1(B). Thus, by an examination of the servo data, the validity of the fundamental assumptions in the evaluation procedure can be checked.

Evaluation with Random-Signal Data

The evaluation technique using random input-response data is similar to the sinusoidal technique. Instead of exciting the servo under test with a sinusoid and measuring the fundamental response component, a Gaussian random input is used, and the mean-square response is determined as a function of the frequency composition and mean-square amplitude of the input. The use of the random-signal procedure is indicated whenever random test signals are more easily obtained or whenever the operating input of the servo is similar to a Gaussian random signal.

The random-signal procedure is developed from a method for calculating the mean-square response of certain
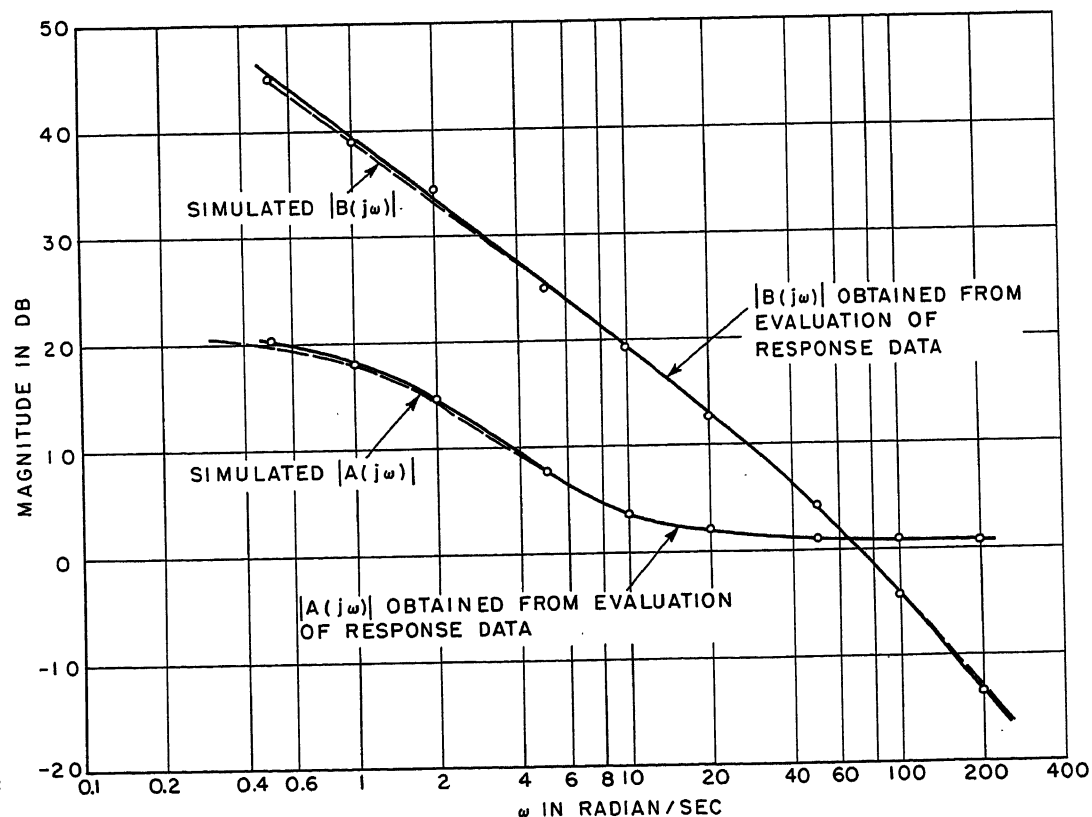
Fig. 5. Transfer functions of linear elements
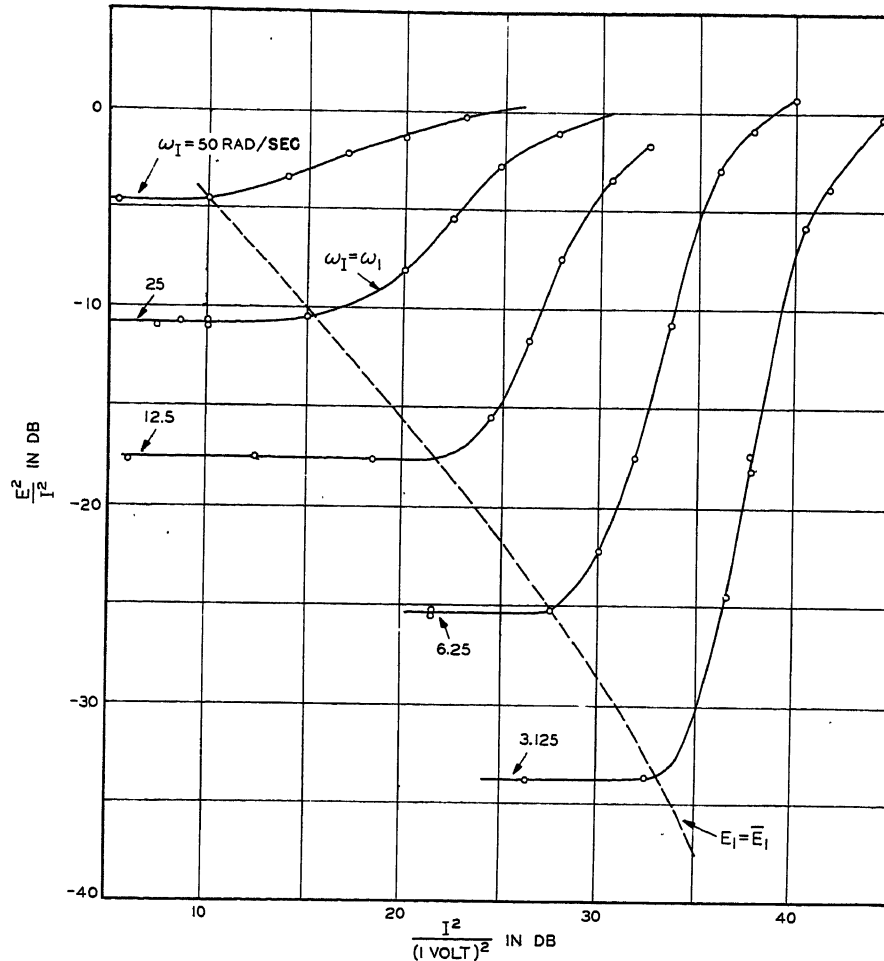
Fig. 6. Random-signal data

types of nonlinear servos presented by Booton.[7] In Booton's method, the nonlinear element is approximated by a quasi-linear element, the parameters of which are functions of the mean-square input of the element. The output of the nonlinear element e_2' is again given by the function

$$e_2' = f(e_1') \qquad (22)$$

where e_1' is the input to the nonlinear element.

The quasi-linear representation of the $f(e_1')$ nonlinear element is a simple gain function[7] $F(E_1)$ which is determined from $f(e_1')$ by the equation

$$F(E_1) = \frac{1}{\sqrt{2\pi E_1^3}} \int_{-\infty}^{+\infty} e_1' f(e_1') \times \exp\left(\frac{-(e_1')^2}{2E_1^2}\right) de_1' \qquad (23)$$

where E_1 is the rms input to the quasi-linear element. Thus, in the quasi-linear system

$$\frac{e_2}{e_1} = F(E_1) \qquad (24)$$

where e_1 and e_2 are the input and output of the quasi-linear element.

With a random input, the mean-square values of the various signals in the servo may be defined by the relations

$$
\begin{aligned}
\overline{i'^2} &= I^2 \\
\overline{e'^2} &= E^2 \\
\overline{e_1'^2} &= E_1^2 \\
\overline{r'^2} &= R^2
\end{aligned}
\qquad (25)
$$

in which the lower-case letters represent instantaneous time functions and the capital letters represent rms values.

If the random input is applied to the quasi-linear system and the basic assumption is made that the quasi-linear system is a sufficiently good approximation to the servo, the corresponding mean-square signals in the two systems may be equated according to the relations

$$
\begin{aligned}
\overline{i^2} &= I^2 \\
\overline{e^2} &= E^2 \\
\overline{e_1^2} &= E_1^2 \\
\overline{r^2} &= R^2
\end{aligned}
\qquad (26)
$$

The data for the servo evaluation consist of the ratio of mean-square error to mean-square input E^2/I^2 measured as a function of I^2 and ω_I. The fre-

quency ω_I is referred to as the cutoff frequency of the input and is defined precisely by the spectrum of the input

$$\Phi_{ii}(\omega) = \Phi_0 \frac{\omega_I^4}{\omega^4 + \omega_I^4} \qquad (27)$$

in which Φ_0 is the zero-frequency spectral density with dimensions of volts squared per radian per second, and the spectral energy is equally divided between negative and positive values of the frequency ω. The spectrum given by equation 27 was chosen because it leads to a simple evaluation procedure. Also, a discussion of one spectrum is sufficient to illustrate the procedure.

A set of data from a typical servo which will be used as an example is shown in Fig. 6. (The same servo was studied sinusoidally in the previous section.) The data shown in Fig. 6 were obtained from a simulation of the servo on an electronic differential analyzer.

Two basic statistical relations are required in the evaluation procedure

$$I^2 = \int_{-\infty}^{+\infty} \Phi_{ii}(\omega) d\omega \qquad (28)$$

$$\Phi_{oo}(\omega) = |H(j\omega)|^2 \Phi_{ii}(\omega) \qquad (29)$$

The first relates the mean-square value I^2 of a random signal to the spectrum $\Phi_{ii}(\omega)$ of the random signal. The second gives the spectrum $\Phi_{oo}(\omega)$ of the output of a linear system in terms of the transfer function $H(j\omega)$ of the system and the spectrum $\Phi_{ii}(\omega)$ of the input.

The mean-square performance of the quasi-linear system may be calculated by the use of equations 28 and 29. The mean-square input may be written as follows

$$I^2 = \Phi_0 \int_{-\infty}^{+\infty} \frac{\omega_I^4}{\omega^4 + \omega_I^4} d\omega \qquad (30)$$

The integral in equation 30 can be evaluated conveniently by the integral tables published in Appendix C of reference 8 to give the relation

$$I^2 = \Phi_0 \frac{\pi \omega_I}{\sqrt{2}} \qquad (31)$$

The mean-square error E^2 and the mean-square input to the nonlinear element E_1^2 may be expressed as

$$E^2 = \Phi_0 \int_{-\infty}^{+\infty} \frac{\omega_I^4}{\omega^4 + \omega_I^4} \times \left| \frac{1}{1 + A(j\omega) B(j\omega) F(E_1)} \right|^2 d\omega \qquad (32)$$

$$E_1^2 = \Phi_0 \int_{-\infty}^{+\infty} \frac{\omega_I^4}{\omega^4 + \omega_I^4} \times \left| \frac{A(j\omega)}{1 + A(j\omega) B(j\omega) F(E_1)} \right|^2 d\omega \qquad (33)$$

By combining equations 31 and 32 and 31 and 33, the following relations are obtained for E^2/I^2 and E_1^2/I^2

$$\frac{\pi\omega_I}{\sqrt{2}}\frac{E^2}{I^2} = \int_{-\infty}^{+\infty} \frac{\omega_I^4}{\omega^4 + \omega_I^4} \times$$
$$\left|\frac{1}{1+A(j\omega)B(j\omega)F(E_1)}\right|^2 d\omega \quad (34)$$

$$\frac{\pi\omega_I}{\sqrt{2}}\frac{E_1^2}{I^2} = \int_{-\infty}^{+\infty} \frac{\omega_I^4}{\omega^4 + \omega_I^4} \times$$
$$\left|\frac{A(j\omega)}{1+A(j\omega)B(j\omega)F(E_1)}\right|^2 d\omega \quad (35)$$

An experimental curve of E^2/I^2 along which E_1 is constant may be used with equations 34 and 35 to evaluate the linear transfer functions $A(j\omega)$ and $B(j\omega)$. A curve of constant E_1, ($E_1 = \bar{E}_1$), shown by the dashed line in Fig. 6, may be located by methods that will be described later. Along this curve E^2/I^2 may be denoted by the relation

$$\frac{E^2}{I^2} = \frac{E^2}{I^2}(\omega_I)\Big|_{\bar{E}_1} \quad (36)$$

and $1/I^2$ may be denoted by the relation

$$\frac{1}{I^2} = \frac{1}{I^2}(\omega_I)\Big|_{\bar{E}_1} \quad (37)$$

both quantities being functions of ω_I. Substitution of equation 36 into 34 yields the integral equation

$$\frac{\pi\omega_I}{\sqrt{2}}\left(\frac{E^2}{I^2}(\omega_I)\Big|_{\bar{E}_1}\right) = \int_{-\infty}^{+\infty} \frac{\omega_I^4}{\omega^4 + \omega_I^4} \times$$
$$\left|\frac{1}{1+A(j\omega)B(j\omega)F(\bar{E}_1)}\right|^2 d\omega \quad (38)$$

which may be solved to determine $A(j\omega)B(j\omega)F(\bar{E}_1)$. The final form of solution as developed in Appendix II may be written

$$\left|\frac{1}{1+A(j\omega)B(j\omega)F(\bar{E}_1)}\right|^2$$
$$= Re\left(\frac{E^2}{I^2}\left(\omega\epsilon^{\frac{j\pi}{4}}\right)\Big|_{\bar{E}_1}\right) +$$
$$Im\left(\frac{E^2}{I^2}\left(\omega\epsilon^{\frac{j\pi}{4}}\right)\Big|_{\bar{E}_1}\right) \quad (39)$$

The solution expressed by equation 39 is written in terms of the function

$$\frac{E^2}{I^2}(\omega_I)\Big|_{\bar{E}_1}$$

in which the argument ω_I is complex. Since the experimental data can be obtained for real values of ω_I only, analytic continuation must be used to extend the data into the complex plane. The details of this continuation are discussed also in Appendix II.

To find $A(j\omega)B(j\omega)F(\bar{E}_1)$ from equation 39, a transfer function $H_1(j\omega)$ must be
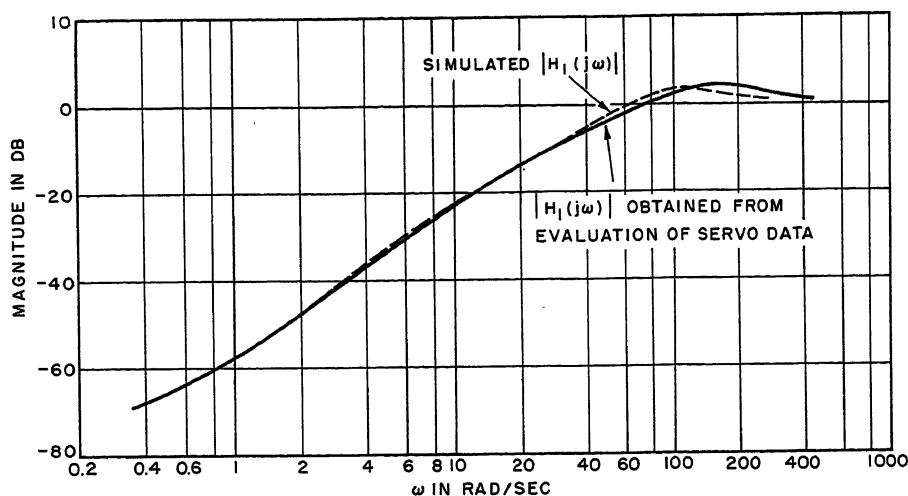
determined which satisfies the relation

$$\left|\frac{1}{1+A(j\omega)B(j\omega)F(\bar{E}_1)}\right| = |H_1(j\omega)| \quad (40)$$

The function $H_1(j\omega)$ can be obtained by using the Bode[6] procedure to associate a phase with the known magnitude function, or by some other standard method. After $H_1(j\omega)$ has been determined, equation 40 can be solved for $A(j\omega)B(j\omega)$ to give the relation

$$A(j\omega)B(j\omega) = \frac{1 - H_1(j\omega)}{F(\bar{E}_1)H_1(j\omega)} \quad (41)$$

The constant $F(\bar{E}_1)$ is one of the arbitrary scale factors that appear in the evaluation procedure.

A solution to obtain $|H_1(j\omega)|$ from the data of Fig. 6 is carried out as an example in Appendix II. The resulting $|H_1(j\omega)|$ shown on Fig. 7 agrees closely with the simulated $|H_1(j\omega)|$. A graphical solution of equation 41 to determine $A(j\omega)B(j\omega)$ with the assumption that $F(\bar{E}_1) = 1$ yields the expression

$$A(j\omega)B(j\omega) = 995\frac{(0.077j\omega+1)}{(0.87j\omega+1)(0.003j\omega+1)} \quad (42)$$

The simulated transfer function is

$$A(j\omega)B(j\omega) = 1000\frac{(0.1j\omega+1)}{(j\omega+1)(0.01j\omega+1)} \quad (43)$$

A comparison of equations 42 and 43 shows that all the time constants have been determined to a reasonable degree of accuracy, with the exception of the 0.01-second time constant. However, since 50 radians per second is the highest frequency for which experimental data were taken, the error is not surprising.

Equation 41 determines the product $A(j\omega)B(j\omega)$. These two transfer functions may be separated by using the data given in equation 37. The substitution of

equation 37 into equation 35 results in the integral equation

$$\frac{\pi\omega_I}{\sqrt{2}}\left(\frac{1}{I^2}(\omega_I)\Big|_{\bar{E}_1}\right) = \int_{-\infty}^{+\infty} \frac{\omega_I^4}{\omega^4 + \omega_I^4} \times$$
$$\left|\frac{A(j\omega)}{1+A(j\omega)B(j\omega)F(\bar{E}_1)}\right|^2 \frac{1}{\bar{E}_1^2} d\omega \quad (44)$$

which may be solved in exactly the same manner as equation 39 to give the relations

$$\frac{1}{\bar{E}_1^2}\left|\frac{A(j\omega)}{1+A(j\omega)B(j\omega)F(\bar{E}_1)}\right|^2$$
$$= Re\left(\frac{1}{I^2}\left(\omega\epsilon^{\frac{j\pi}{4}}\right)\Big|_{\bar{E}_1}\right) +$$
$$Im\left(\frac{1}{I^2}\left(\omega\epsilon^{\frac{j\pi}{4}}\right)\Big|_{\bar{E}_1}\right) \quad (45)$$
$$= |H_2(j\omega)|^2 \quad (46)$$

The combination of equations 40 and 46 yields the relations

$$A(j\omega) = \bar{E}_1\frac{H_2(j\omega)}{H_1(j\omega)} \quad (47)$$

$$B(j\omega) = \frac{1 - H_1(j\omega)}{\bar{E}_1 F(\bar{E}_1)H_2(j\omega)} \quad (48)$$

for $A(j\omega)$ and $B(j\omega)$, thus completing the evaluation of the linear parts of the quasi-linear system. The quantity $\bar{E}_1$ in these two relations is the second scale-factor constant.

A solution of equations 45, 46, and 47 to obtain $|A(j\omega)|$ from the data of Fig. 6 results in the function shown in Fig. 8. The evaluated $|A(j\omega)|$ differs considerably from the simulated $|A(j\omega)|$, but the evaluated transfer function shows definitely which terms in equation 42 should be associated with $A(j\omega)$. Therefore, $A(j\omega)$ and $B(j\omega)$ can be separated and written

$$A(j\omega) = 9.95\frac{0.077j\omega+1}{0.87j\omega+1} \quad (49)$$

Fig. 7. Evaluation of $|H_1(j\omega)|$
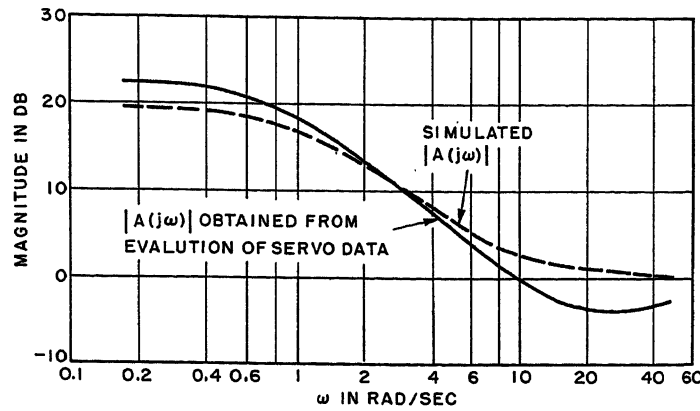
Fig. 8 (left). Evaluation of $|A(j\omega)|$
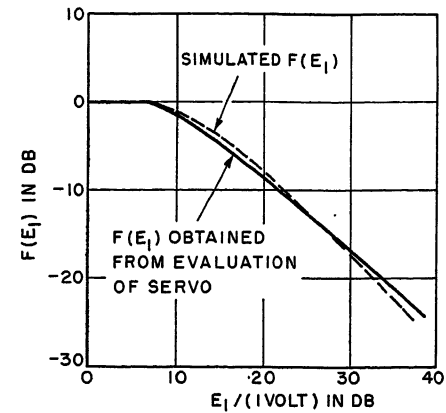
Fig. 9 (right). Quasi-linear gain function from random-signal data

$$B(j\omega) = \frac{100}{j\omega(0.003j\omega+1)} \quad (50)$$

where the second scale-factor constant E_1 is chosen to give $B(j\omega)$ a low-frequency gain of $100/s$.

To evaluate the gain function of the quasi-linear element, data from one of the curves of constant ω are used. The curve in Fig. 6 along which

$$\omega_I = \omega_1 \quad (51)$$

is used, and the value of E^2/I^2 along this curve is denoted

$$\frac{E^2}{I^2} = \frac{E^2}{I^2}(I^2)\Big|_{\omega_1} \quad (52)$$

The substitution of ω_1 and the values of $A(j\omega)$ and $B(j\omega)$, obtained from equations 47 and 48, into equation 34 and the evaluation of the integral yield E^2/I^2 as a function of $F(E_1)$. The inverse function $F(E_1)$ as a function of E^2/I^2 is denoted $S_1(E^2/I^2)$

$$F(E_1) = S_1(E^2/I^2) \quad (53)$$

Equations 52 and 53 may be formally solved to yield the relation

$$F(E_1) = S_1\left(\frac{E^2}{I^2}(I^2)\Big|_{\omega_1}\right) \quad (54)$$

for $F(E_1)$ in terms of the experimental data given by equation 52. The actual solution is conveniently carried out in the tabular form illustrated in Table III. The values of $F(E_1)$ and E^2/I^2 obtained from equation 53 are listed in the first two columns. The experimental data (equation 52) are used to determine the

Table III. Evaluation of F(E_1) from Random Data

F(E_1), Db	E²/I², Db	I²/(1 Volt)², Db	E₁²/I², Db	E₁/1 Volt, Db
0	−9.8	17.2	−10.2	7.1
− 0.9	−9.1	18.8	− 9.5	9.2
− 1.9	−8.3	19.8	− 8.7	11.1
− 6.0	−5.4	22.5	− 5.6	16.9
−14.0	−1.2	27.3	− 0.8	26.6
−20.0	0.5	31.8	1.8	33.6

values of I^2 in the third column that correspond to the values of E^2/I^2 in the second column. The evaluation is completed by substituting $A(j\omega)$, $B(j\omega)$, and ω_1 into equation 35 to determine E_1^2/I^2 as a function of $F(E_1)$ which is tabulated in the fourth column. The input E_1 then can be calculated as the square root of the product of the third and fourth columns; the results of this computation are tabulated in the last column. Good agreement is obtained between $F(E_1)$ and the simulated quasi-linear gain function, both of which are shown in Fig. 9. The nonlinear response function obtained from $F(E_1)$ by the methods of Appendix I is shown in Fig. 10 together with the simulated response function. These functions also agree closely.

The determination of $F(E_1)$ completes the evaluation of the quasi-linear system, except for the location of a line of constant E_1. This location is not as simple as in the sinusoidal case because the random-data curves are not similar to one another. However, several ways of locating a line of constant E_1 will be mentioned.

In some cases the effect of a distinctive characteristic of the nonlinear element can be noticed on each curve of E^2/I^2 versus I^2. These effects serve as a simple means of locating the line of constant E_1. In the example of Fig. 6, the sharp knee in the nonlinear response curve produces a prominent change in the curvature of each E^2/I^2 curve, and the $E_1 = \bar{E}_1$ line is drawn through the points at which the curvature changes.

Some servos exhibit a noticeable region of linear operation which may be used to locate a curve of constant E_1. In a region of linear operation, $F(E_1)$ is constant, and an area of horizontal E^2/I^2 curves exists, as illustrated in Fig. 6, because

$$\frac{dF(E_1)}{dE_1} = 0 \quad (55)$$

In this linear region, E^2/I^2 can be

evaluated as a function of ω_I and these data used to solve equation 38 to determine $A(j\omega)B(j\omega)$. A solution of equation 38 to obtain $A(j\omega)B(j\omega)$ is possible because $F(E_1)$ is constant in the linear region. However, E_1 is not constant in this region, and thus a solution of equation 44 to separate $A(j\omega)$ and $B(j\omega)$ cannot be performed directly. Instead, the value of $A(j\omega)B(j\omega)$ is substituted into equation 34 along with some constant value of $F(E_1)$ to yield E^2/I^2 as a function of ω_I. The constant value of $F(E_1)$ must be chosen in a nonlinear region. The function E^2/I^2 may be combined with the experimental data to locate a line along which E_1 is constant.

If no region of linear operation and no distinctive characteristics of the nonlinear element exist, an iteration process is necessary to locate a line of constant E_1. The iteration is begun by guessing the location of a line. Data taken from this line are used to evaluate $A(j\omega)B(j\omega)$ and thus to locate a second line of constant E_1. Data from the second line are used to re-evaluate $A(j\omega)B(j\omega)$ and thus to obtain a new estimate for the location of the original line. The process of locating alternately the two lines of constant E_1 is repeated until the locations become fixed.
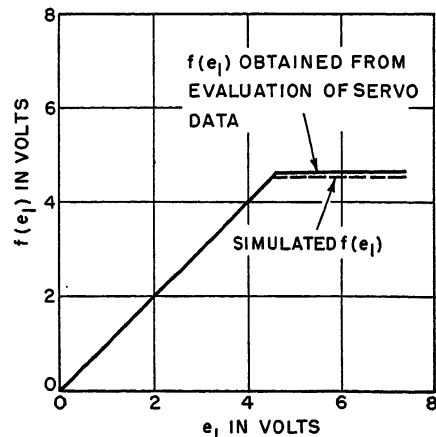
Fig. 10. Response function from random-signal data

Conclusions

The sinusoidal and random-signal evaluation methods constitute an example of the application of quasi-linear methods for the evaluation of nonlinear systems. These methods are used in combination with a particular assumption concerning the form of the system to yield a specific procedure for evaluating a limited, but important, class of non-linear servos. By this means, the soundness and usefulness of the quasi-linear approach to evaluation problems are demonstrated. This demonstration is the most significant point established by the research. The success of quasi-linear methods for the evaluation of one particular form of servo indicates that future work could be directed profitably toward extending the procedure to other forms of servosystems.

The sinusoidal procedure is a direct extension of existing techniques and, as such, is readily comprehended. Commercial equipment for measuring phase and magnitude may be used to take the required data. Thus, the procedure is applicable to problems of immediate and practical interest.

The practical importance of the random-signal procedure is not as great as that of the sinusoidal procedure because the random-signal techniques are more complex, both computationally and experimentally. However, the random-signal studies are of theoretical interest and may develop into an important evaluation technique.

Appendix I. Evaluation of the Nonlinear Response Function

Methods for evaluating the quasi-linear system that approximates a nonlinear servo have been presented. To complete the evaluation of the servomechanism, the response function of the nonlinear element must be determined from the gain function of the quasi-linear element. In the sinusoidal case, the quasi-linear gain function and the response function are related by equation 2, which is repeated here

$$F(E_1) = \frac{2}{\pi E_1} \int_{-\pi/2}^{+\pi/2} f(E_1 \sin \theta) \sin \theta \, d\theta \quad (56)$$

If the quasi-linear gain function $F(E_1)$ is given and the response function $f(e_1')$ is unknown, equation 56 is an integral equation for which an approximate solution can be obtained in a number of ways. A graphical procedure, which is both simple and satisfactory, approximates $F(E_1)$ with the sum

$$F(E_1) \cong \sum_{n=1}^{N} \beta_n F_n(E_1) \quad (57)$$

The approximation is carried out by choosing appropriate β_n coefficients and $F_n(E_1)$ functions. The $F_n(E_1)$ functions are obtained by substituting known approximating functions $f_n(e_1')$ into equation 56 and evaluating the resulting integral

$$F_n(E_1) = \frac{2}{\pi E_1} \int_{-\pi/2}^{+\pi/2} f_n(E_1 \sin \theta) \sin \theta \, d\theta \quad (58)$$

The response function $f(e_1')$ then is given by the relation

$$f(e_1') \cong \sum_{n=1}^{N} \beta_n f_n(e_1') \quad (59)$$

The semi-infinite slope function shown in Fig. 11 is a convenient approximating function $f_n(e_1')$. If this function is substituted in equation 58, the resulting integral may be written

$$F_n(E_1) = F^*\left(\frac{E_1}{\alpha_n}\right) = \frac{4\alpha_n}{\pi E_1} \int_{\sin^{-1}\left(\frac{\alpha_d}{E_1}\right)}^{\pi/2} \times \left(\frac{E_1}{\alpha_n} \sin \theta - 1\right) \sin \theta d \, \theta \quad (60)$$

and, as is apparent from equation 60, all the $F_n(E_1)$ functions may be written as a single function $F^*(E_1/\alpha_n)$ of the argument E_1/α_n. Thus $F(E_1)$ may be rewritten as the summation

$$F(E_1) \cong \sum_{n=1}^{N} \beta_n F^*\left(\frac{E_1}{\alpha_n}\right) \quad (61)$$

and the solution of the integral equation may be carried out by choosing values of α_n and β_n to approximate $F(E_1)$.

The α_n and β_n coefficients may be chosen by plotting ln $F(E_1)$ versus ln E_1 and ln $F^*(E_1/\alpha_n)$ versus ln E_1/α_n. These two plots may be superimposed over a light box and shifted up and down and right and left to determine the best fit and thus to determine α_1 and β_1. The function

$$F(E_1) - \beta_1 F^*\left(\frac{E_1}{\alpha_1}\right)$$

then can be plotted and the fitting process repeated to determine α_2 and β_2. The fitting process may be repeated as many times as is necessary to obtain a good fit · to $F(E_1)$. Since the process yields a line-segment approximation to $f(e_1')$, any $f(e_1')$ function can be fitted by using a sufficient number of line segments.

The quasi-linear gain function determined in the example and shown in Fig. 3 may be approximated by two terms in the sum given by equation 61. This sum may be written

$$F(E_1) \cong F^*\left(\frac{E_1}{0}\right) - F^*\left(\frac{E_1}{10}\right) \quad (62)$$

The functions $F^*(E_1/0)$ and $F^*(E_1/10)$ are also shown in Fig. 3. The difference of these functions is $F(E_1)$, although this difference is not directly apparent in the figure owing to the logarithmic co-ordinates.

The nonlinear response function $f(e_1')$ obtained by substituting the values of α_n and β_n implied by equation 62 into equation 59 is plotted in Fig. 4. Within the accuracy of the plot, this response function is identical to the simulated response function.

Appendix II. Solution of Transfer-Function Integral Equations

In the random-signal evaluation, solutions of integral equations 38 and 44 are required. These equations are of the form

$$\frac{\pi \omega_I}{\sqrt{2}}[R(\omega_I)] = \int_{-\infty}^{+\infty} \frac{\omega_I}{\omega^4 + \omega_I^4} |H(j\omega)|^2 d\omega \quad (63)$$

in which the unknown function $H(j\omega)$ is a transfer function of the servo and $R(\omega_I)$ is an experimentally determined ratio of mean-square voltages.

The form of the kernel of equation 63 makes possible the following solution by Laplace transform methods. Owing to the evenness of the integrand, equation 63 may be rewritten in the form

$$\frac{\pi \omega_I}{2\sqrt{2}}[R(\omega_I)] = \int_0^\infty \frac{\omega_I^4}{\omega^4 + \omega_I^4} |H(j\omega)|^2 d\omega \quad (64)$$

After changing the variables, $\omega_n = \epsilon^{-\tau}$ and $\omega = \epsilon^{-\sigma}$, equation 64 becomes

$$\frac{\pi}{2\sqrt{2}} \epsilon^{-\tau} R(\epsilon^{-\tau})$$

$$= \int_{-\infty}^{+\infty} \frac{1}{1 + \epsilon^{4(\tau - \sigma)}} |H(j\epsilon^{-\sigma})|^2 \epsilon^{-\sigma} d\sigma \quad (65)$$

The kernel of equation 65 is a function of the difference $\tau - \sigma$ of the two variables. Consequently, this integral equation can be solved with a 2-sided Laplace transform by means of the convolution theorem.[9]

By use of this theorem, the solution to equation 65 may be written

$$|H(j\epsilon^{-\sigma})|^2 \epsilon^{-\sigma} = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} \frac{C(s)}{D(s)} \epsilon^{s\sigma} ds \quad (66)$$

in which $C(s)$ is the transform of the left-hand side of equation 65

$$C(s) = \int_{-\infty}^{+\infty} \frac{\pi}{2\sqrt{2}} \epsilon^{-\tau} R(\epsilon^{-\tau}) \epsilon^{-\tau s} d\tau \quad (67)$$

and $D(s)$ is the transform of the kernel of equation 65

$$D(s) = \int_{-\infty}^{+\infty} \frac{1}{1 + \epsilon^{4\sigma}} \epsilon^{-\sigma s} d\sigma \quad (68)$$

If the transforms $C(s)$ and $D(s)$ have a common strip of convergence, equation 66 is a straightforward method for obtaining a solution to the integral equation 63 since the integrals in Eqs. 66, 67, and 68 can be calculated numerically. However, because the numerical procedure is tedious, a simpler process involving the analytic continuation of $R(\omega_I)$ to complex values of ω_I is developed here.

The integral in equation 68 can be evaluated by contour integration to give the relation

$$D(s) = -\frac{\pi}{4 \sin \frac{\pi}{4} s} \quad (69)$$

The quantity $C(s)/D(s)$ may then be expressed

$$\frac{C(s)}{D(s)} = -\frac{2}{j\pi}\left(\epsilon^{\frac{j\pi}{4}s} - \epsilon^{-\frac{j\pi}{4}s}\right)C(s) \quad (70)$$

Substitution of equation 70 into 66 yields an integral which may be evaluated in terms of $R(\omega_I)$ to yield the relation

$$|H(j\epsilon^{-\sigma})|^2\epsilon^{-\sigma} = \frac{j}{\sqrt{2}}\left[\epsilon^{-\sigma}\epsilon^{-\frac{j\pi}{4}}\times\right.$$
$$\left. R\left(\epsilon^{-\sigma}\epsilon^{-\frac{j\pi}{4}}\right)-\epsilon^{-\sigma}\epsilon^{\frac{j\pi}{4}}R\left(\epsilon^{-\sigma}\epsilon^{\frac{j\pi}{4}}\right)\right] \quad (71)$$

After the change of variable $\omega=\epsilon^{-\sigma}$, equation 71 may be reduced to the form

$$|H(j\omega)|^2 = Re\left[R\left(\omega\epsilon^{\frac{j\pi}{4}}\right)\right]+Im\left[R\left(\omega\epsilon^{\frac{j\pi}{4}}\right)\right] \quad (72)$$

Equation 72 expressed $|H(j\omega)|^2$ as a function of the real and imaginary parts of $R(\omega\epsilon^{j\pi/4})$. Experimental data taken from the servo yield values of $R(\omega_I)$ for only real values of ω_I. Consequently, a process of analytic continuation must be used to extend the experimental data to complex values of ω_I. This analytic continuation can be carried out conveniently by approximating $R(\omega_I)$ by an analytic function. A definite form for this function is indicated because $H(j\omega)$ is a rational function of $j\omega$, and therefore $R(\omega_I)$ is a rational function of ω_I for positive values of ω_I, as shown by the integral tables in Appendix C of reference 8. Consequently, the form

$$R(\omega_I) = \frac{b_0+b_1\omega_I+b_2\omega_I{}^2+\ldots+b_n\omega_I{}^n}{a_0+a_I\omega_I+a_2\omega_I{}^2+\ldots+a_m\omega_I{}^m} \quad (73)$$

is appropriate. The a and b coefficients are chosen to make $R(\omega_I)$ fit the experimental data. Equation 73 can be expanded into a set of linear algebraic equations which may be solved to obtain a point fit. As an example, a solution to equation 38 using the data from Fig. 6 will be carried out. The ratio E^2/I^2 for constant E_1 is plotted as a function of ω_I in Fig. 12. The process of fitting $R(\omega_I)$ to this function is begun by rewriting equation 73 in the form

$$\omega_I R(\omega_I)a_1+\omega_I{}^2R(\omega_I)a_2+\omega_I{}^3R(\omega_I)a_3+$$
$$\omega_I{}^4R(\omega_I)a_4+\omega_I{}^5R(\omega_I)a_5-\omega_I{}^3b_3-$$
$$\omega_I{}^4b_4-\omega_I{}^5b_5=\omega_I{}^6[1-R(\omega_I)]+$$
$$10^{-6}\omega_I{}^2-R(\omega_I) \quad (74)$$

Incorporated in equation 74 is the assumed information that the forward-loop transfer function of the servo approaches zero at high frequencies and approaches $1000/s$ at low frequencies. Thus, from equation 34

$$\frac{E^2}{I^2}(\omega_I)\to 1, \text{ as } \omega_I\to\infty$$

$$\frac{E^2}{I^2}(\omega_I)\to\frac{\omega_I{}^2}{10^6}, \text{ as } \omega_I\to 0$$

The values of ω_I and $R(\omega_I)$ from the circled data points in Fig. 12 are substituted into equation 74 to obtain eight relations which may be written as the matrix equation

Fig. 11 (right). Approximating function

Fig. 12 (below). Data from curve of constant E_1
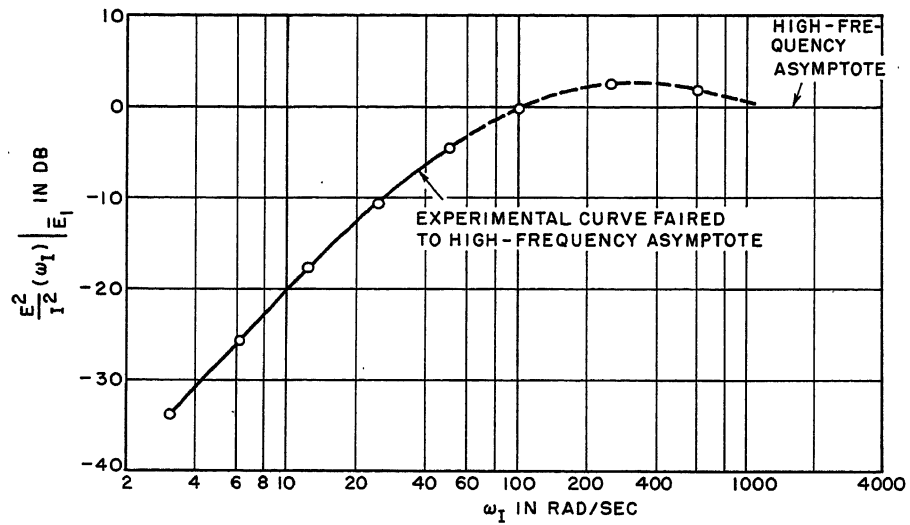
Several multiplying factors have been introduced into the rows and columns of equation 75 to reduce the range of the matrix coefficients. A solution, obtained by using the Crout[10] method and a desk calculator, is

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} 3.27\times 10^{-2} \\ 3.56\times 10^{-4} \\ -49.6\ \times 10^{-6} \\ 28.0\ \times 10^{-8} \\ -24.1\ \times 10^{-10} \\ 15.8\ \times 10^{-6} \\ -28.2\ \times 10^{-8} \\ -24.3\ \times 10^{-10} \end{bmatrix} \quad (76)$$

Substitution of the analytic function in equation 73 with the coefficients from equation 76 into equation 72 yields the value of $|H_1(j\omega)|$ plotted in Fig. 7, thus completing the solution of equation 38.

References

1. MEASUREMENT OF SOME NONLINEARITIES IN SERVOMECHANISMS, Dietrich K. Gehmlich, M. E.

Van Valkenburg. *AIEE Transactions*, vol. 73, pt. II, Nov. 1954, pp. 232–35.

2. A FREQUENCY RESPONSE METHOD FOR ANALYZING AND SYNTHESIZING CONTACTOR SERVOMECHANISMS, Ralph J. Kochenburger. *Ibid.*, vol. 69, pt. I, 1950, pp. 270–84.

3. SINUSOIDAL ANALYSIS OF FEEDBACK-CONTROL SYSTEMS CONTAINING NONLINEAR ELEMENTS, E. Calvin Johnson. *Ibid.*, vol. 71, pt. II, July 1952, pp. 169–81.

4. THEORY OF SERVOMECHANISMS (book), H. M. James, N. B. Nichols, R. S. Phillips. McGraw-Hill Book Company, Inc., New York, N. Y., vol. 25, Radiation Laboratory Series, 1947, pp. 179–86.

5. A TECHNIQUE FOR THE EVALUATION OF NONLINEAR SERVOMECHANISMS, M. V. Mathews. *Sc.D. Thesis*, Massachusetts Institute of Technology, Cambridge, Mass., 1954, pp. 13–14.

6. NETWORK ANALYSIS AND FEEDBACK AMPLIFIER DESIGN (book), H. W. Bode. D. Van Nostrand Company, Inc., New York, N. Y., 1945, chap. XV.

7. NONLINEAR CONTROL SYSTEMS WITH STATISTICAL INPUTS, R. C. Booton, Jr. *Dynamic Analysis and Control Laboratory Report No. 61*, Massachusetts Institute of Technology, Cambridge, Mass., Mar. 1, 1952.

8. NONLINEAR SERVOMECHANISMS WITH RANDOM INPUTS, R. C. Booton, Jr., M. V. Mathews, W. W. Seifert. *Dynamic Analysis and Control Laboratory Report No. 70, Ibid.*, Aug. 20, 1953.

9. OPERATIONAL CALCULUS BASED ON THE TWO-

$$\begin{bmatrix} 13.1 & 41 & 1.28 & 0.4 & 0.0125 & 3,050 & 954 & 29.8 \\ 17.8 & 111 & 6.96 & 4.35 & 0.272 & 2,240 & 1,520 & 95.4 \\ 21.2 & 265 & 33.2 & 41.5 & 5.19 & 1,950 & 2,440 & 305 \\ 210 & 5,250 & 1,310 & 3,280 & 820 & 15,600 & 39,000 & 9,770 \\ 17.5 & 875 & 437 & 2,180 & 1,090 & 1,250 & 6,250 & 3,120 \\ 97 & 9,700 & 9,700 & 97,000 & 97,000 & 10,000 & 100,000 & 100,000 \\ 4.2 & 1,060 & 2,656 & 66,400 & 166,000 & 1,560 & 39,100 & 97,700 \\ 0.918 & 551 & 3,300 & 198,000 & 1,189,728 & 2,160 & 129,600 & 778,000 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ 10^2 a_3 \\ 10^3 a_4 \\ 10^5 a_5 \\ -10^2 b_3 \\ -10^3 b_4 \\ -10^5 b_5 \end{bmatrix} = \begin{bmatrix} -4.1 \\ -2.81 \\ -1.68 \\ -8.31 \\ -0.387 \\ -0.93 \\ -1.72 \\ -24.7 \end{bmatrix} \quad (75)$$

SIDED LAPLACE TRANSFORM (book), B. van der Pol, H. Bremmer. Cambridge University Press, London, England, 1950, p. 39.

10. A SHORT METHOD FOR EVALUATING DETERMINANTS AND SOLVING SYSTEMS OF LINEAR EQUATIONS WITH REAL OR COMPLEX COEFFICIENTS, P. D. Crout. *AIEE Transactions*, vol. 60, 1941, pp. 1235–41.

————— ◆ —————

Discussion

George C. Newton, Jr. (Massachusetts Institute of Technology, Cambridge, Mass.): The author is to be congratulated for his study of ways of measuring nonlinearities in control systems by means of input, output, and error measurements only. The revelation in the section entitled "Evaluation with Random-Signal Data" that a nonlinearity can be assessed by means of a random signal is especially interesting from a theoretical point of view. One who is not familiar with these procedures might fear that the validity of a quasi-linear model would have more of a tendency to break down in the random-signal case than in the sinusoidal-signal case because of distortion in the probability density functions following the clipping action of the nonlinear element. More discussion of this point would be appreciated. At the same time the author may wish to comment on the allowable limits for the departure of the test signal from the normal distribution which can be allowed and still have a workable procedure.

As Mr. Mathews remarks, the sinusoidal test procedure described in the section entitled "Evaluation with Sinusoidal Data" is probably the more useful technique for assessing nonlinearities. The procedure should be particularly useful for interpreting the frequency-response data obtained from control systems which, although designed on a linear basis, do exhibit appreciable nonlinearity. Since saturation is one of the most frequently encountered forms of nonlinearity in systems designed on a linear basis, the author's procedure ought to be particularly pertinent.

It should be noted that his procedure is based on a single-valued nonlinear function. Thus saturation accompanied by appreciable hysteresis may not be able to be treated by the author's method. In any event, we should have a simple test to determine if the nonlinearity of the system under test may be approximated as being single valued. One simple check is to compare the phase of the output relative to the error under linear operating conditions with the phase under nonlinear conditions. If the nonlinearity is single valued the phase should not be affected by changing from linear to nonlinear operation.

Assuming a single-valued nonlinearity, so that the author's procedures can apply, the experimenter may be faced with a system which exhibits a jump resonance.[1] Under conditions of jump resonance it may be impossible to obtain input over-error versus error magnitude data in the vicinity of the resonant frequency. This may not be a serious matter if good data can be obtained at frequencies well below or well above resonance. However, there are many situations where it is difficult to make accurate error measurements (when the system is operating closed loop) at frequencies very far below resonance and also in which it is difficult to make output measurements at frequencies appreciably above resonance. In these situations does the author have a technique for circumventing the difficulty caused by jump resonance?

REFERENCE

1. SOME SATURATION PHENOMENA IN SERVO-MECHANISMS WITH EMPHASIS ON THE TACHOMETER STABILIZED SYSTEM, E. Levinson. *AIEE Transactions*, vol. 72, pt. II, Mar. 1953, pp. 1–9.
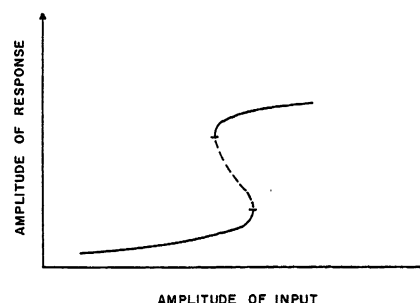
Fig. 13. Jump-resonance phenomenon

M. V. Mathews: Professor Newton's comments are an important addition to the paper, particularly those concerning the application of the evaluation procedure to essentially linear systems with nonlinear aberrations. The evaluation technique was developed mainly to treat such systems. The test suggested by Professor Newton for single-valued nonlinear functions is very useful.

Present experimental data indicate that random-signal testing procedures can be applied to essentially the same class of systems that may be tested with sinusoidal methods. Further information about the accuracy of random-signal techniques has been published in references 7 and 8 of the paper.

As Professor Newton points out, if a jump-resonance phenomenon is present, the response data cannot be obtained over the entire range of servomechanism operation. An example of jump resonance is shown in Fig. 13. Experimental data can be obtained only for the stable parts of the input-response curve shown as solid lines. The dashed curve cannot be obtained. However, in many cases the solid curves can be extrapolated to determine the dashed curve to within the desired accuracy.

Design and Application of a Peak Voltage Detector to Industrial Control Systems

LLOYD W. ALLEN
ASSOCIATE MEMBER AIEE

MANY industrial control, or servo, systems are characterized by their high power level, low cost, and moderate performance when compared with the high-performance servo systems that are used extensively in military applications. This paper describes the control portion of such an industrial type of servo system in some detail. Application of the control portion, or peak voltage detector, to a complete servo system is described briefly.

Three important characteristics of the peak voltage detector are: the nature of the input signals; the nature of the output signals; and the cost of component parts. One input signal is an alternating voltage and one is a direct voltage; either or both inputs may be variable. The detector is capable of indicating a small difference between large magnitudes of the input signals. Output signals may be either of the continuous type or of the discontinuous off-on type; however, the detector is peculiarly adaptable to a discontinuous relay output. For a moderate cost of component parts, the detector with relay output can furnish a stable, reliable output with good sensitivity and large power amplification.

The relay type of peak voltage detector, as defined here, consists of a comparator which compares the peak magni-

tudes of the two input signals, an a-c amplifier which amplifies the resultant signal; a rectifier circuit which converts the amplified signal to direct current; and a sensitive relay which gives an output signal by contact closure. Most applications require two or more such units, operating from different levels of the input signals, to effect the desired output signals. Such units are called "channels" here; a channel is defined as an output relay with its associated circuit.

Fig. 1 shows a single-channel detector in a simplified form. When the peak magnitude of e_{ac} exceeds the magnitude of e_{dc}, $X1$ allows current to pass through $R3$, resulting in a voltage e_{g1} in the form of a series of pulses shaped like the peak of

e_{ac}. After e_{g1} is amplified, $X2$ and its associated filter rectify the amplifier output and the resulting d-c signal e_{g2} operates the output relay.

It will be seen that e_{g2} is a stable amplified signal that is a function of the difference between the two input signals in the limited range where the peak magnitude of e_{ac} is greater than the magnitude of e_{dc} but not sufficiently greater to result in overloading of the amplifier. This relation indicates the possibility of using the continuous variation of e_{g2} as a d-c output voltage signal in place of the discontinuous relay output shown in Fig. 1. A factor that argues in favor of the continuous output is the stability and reliability of e_{g2} when compared with con-

ventional d-c difference amplifier circuits that could replace the peak voltage detector. The increased stability and reliability of e_{g2} is achieved because the a-c nature of e_{g1} allows the use of an inherently stable a-c amplifier without the need for electromechanical chopper stabilization that is commonly used with d-c amplifiers. Factors that argue against the use of e_{g2} as a continuously variable output include the presence of ripple in e_{g2} because of rectification, and the effect of the wave form of e_{ac} on the linearity of the response of e_{g2}. For example, if e_{ac} is a square wave, then e_{g2} may easily be made to vary linearly with the variation of the difference $(e_{ac}-e_{dc})$; however, in the more usual case where e_{ac} is a sine wave, e_{g2} will not easily be made to vary linearly with a variation of $(e_{ac}-e_{dc})$. Further, the power level of e_{g2} is considerably below the power level of a relay-type output for equal component costs. For special applications, then, it appears that the use of e_{g2} as a continuously varying output may be justifiable; but for lowest cost per unit of controlled power the relay-type output appears preferable.

A simpler version of the peak voltage detector can be made by replacing the resistor $R3$ with a sensitive relay, thus eliminating the amplifier and rectifier circuits of Fig. 1. A comparison of this simpler circuit with the circuit of Fig. 1 reveals three important advantages of the amplifier circuit:

1. It presents a higher impedance to the input signals than does the simpler circuit.

2. It may be designed to handle larger overloads.

3. Most important of all, it can be made to reduce effectively the inactive zone of the relay (defined as the difference between the amount of signal voltage e_{g1} required to pick up the relay and the amount of e_{g1} required to drop out the relay) to an arbitrarily small value by the insertion of a sufficiently large amount of voltage gain between e_{g1} and the voltage across the coil of the relay.

The single-channel detector of Fig. 1 is limited to use in control systems where a single off-on response is sufficient to actuate the controlled quantity. This limitation can be removed by the use of a 2-channel detector such as that shown in Fig. 2. In the 2-channel detector, channel 2 operates on a higher signal level than does channel 1. The difference in operating levels is determined by the magnitude of e_z, a small direct voltage that is inserted between $X12$ and $R6$. The output contact of channel 1 is normally closed (relay de-energized) while the output contact of channel 2 is normally open (relay de-energized). When used in a control system, channel 1 operates on the

Fig. 1 (left). Single - channel detector

Fig. 2 (below). Two-channel detector

CH 4
CH 3
CH 2
CH 1
VOLTS
DEAD ZONE
SLOW SPEED ZONE
HIGH SPEED ZONE
e_{ac}

R40 — CH 4
R30 — CH 3
R20 — CH 2
R10 — CH 1

e_{g1} — AMPLI-FIER — RECTI-FIER — e_{g2}

R1
AC INPUT
e_{ac}
R2
R3
R5
e_{dc}
R4
D C INPUT

AC VARIABLE INPUT REFERENCE

D C INPUT

OUTPUT

Fig. 3 (left). Four-channel detector

Fig. 4 (above). Manually variable reference
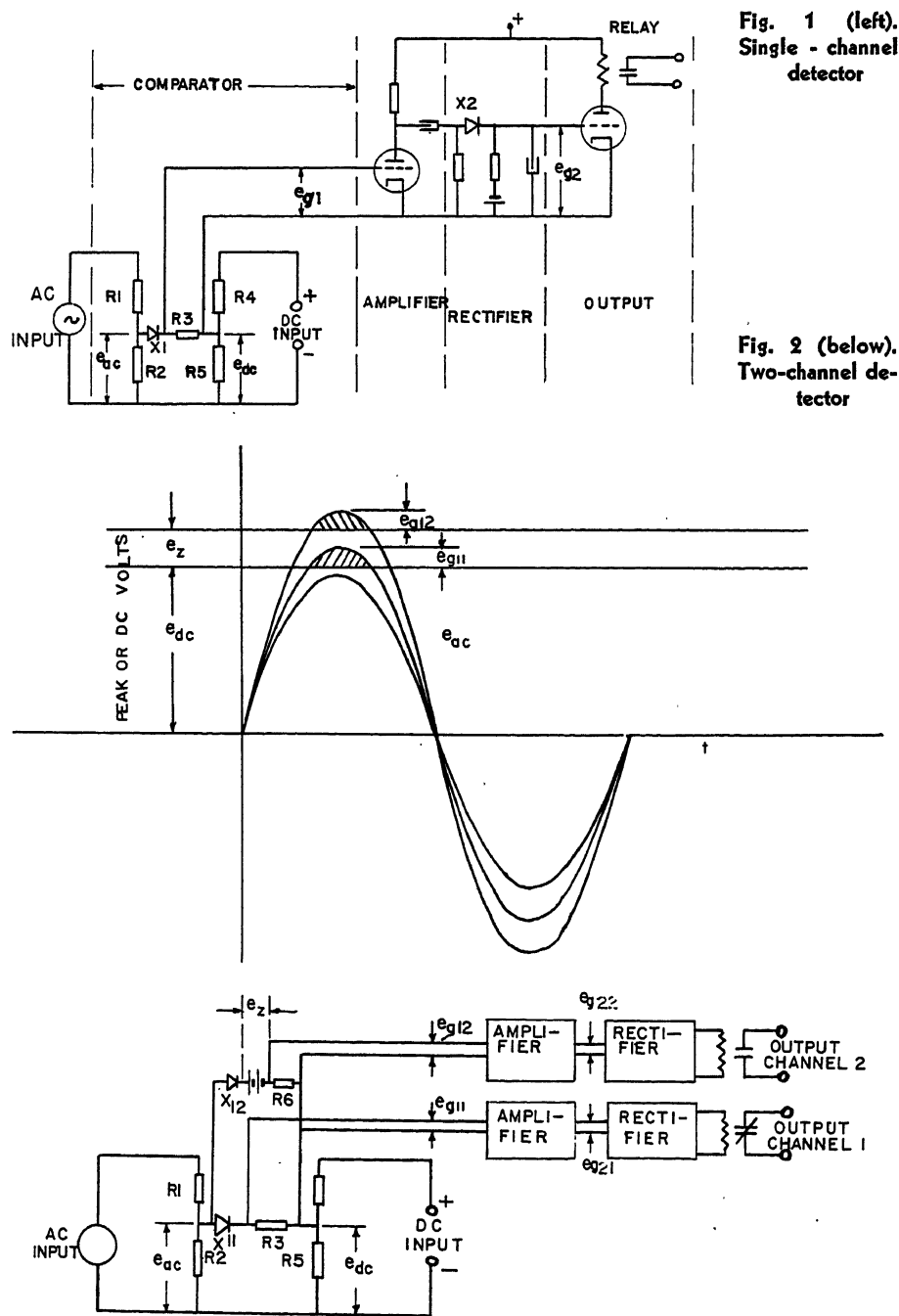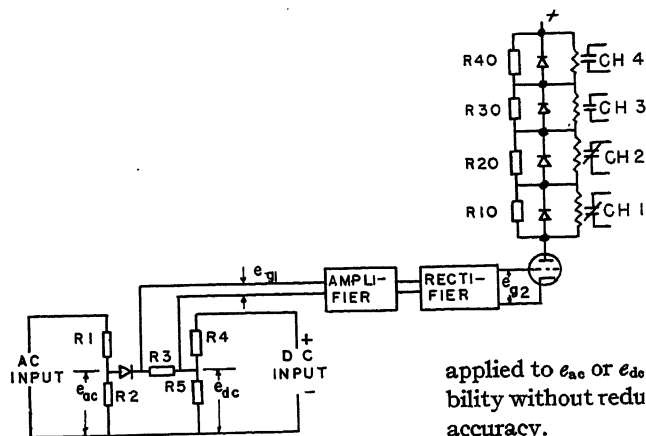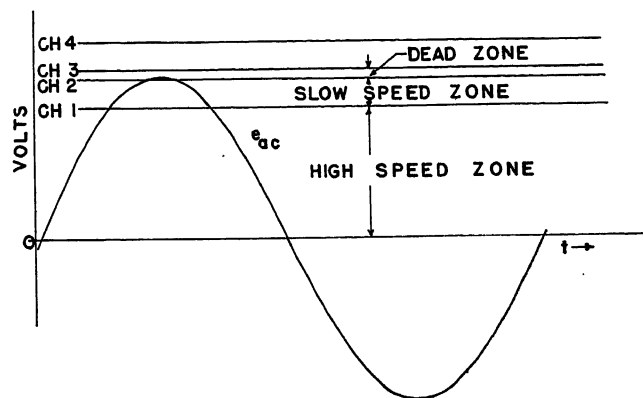
controlled quantity to raise the variable e_{ac} (or lower e_{dc}) as long as its contact remains closed; when the peak value of e_{ac} exceeds e_{dc}, the channel-1 contacts open and the controlled quantity assumes its steady-state condition after an overshoot whose magnitude it determined by the system gain and time constants. If an external disturbance then causes e_{ac} to exceed $(e_{dc}+e_z)$, the channel-2 contacts close and lower e_{ac} until the peak of e_{ac} is again within the dead zone. The dead zone is defined as the region of steady-state operation of the controlled quantity, and is equal in magnitude to e_z. e_z is also the maximum steady-state error signal.

The sensitivity of the detector is a factor in the system gain. The usual definition of sensitivity in a continuous type of controller as output/input must be modified in a discontinuous type of control such as this. Since it is seen that the smaller e_z is in relation to e_{dc} the more sensitive is the corrective action of the peak voltage detector in response to changes in $(e_{ac}-e_{dc})$, sensitivity's here defined as e_{dc}/e_z. For given system time constants, system stability can readily be controlled by adjustment of the value of e_z; however, if stability is increased by an increase of e_z, the system error is also increased proportionately. Conventional feedback stabilization methods[1,2] may be

applied to e_{ac} or e_{dc} to increase system stability without reducing the desired system accuracy.

Another method of increasing system performance is attained by means of a 4-channel detector, such as that shown in Fig. 3. The design of the detector shown in Fig. 3, when compared with the design of Fig. 2, offers several advantages and one important disadvantage. The advantages of the circuit of Fig. 3 are: that less components are required; that more uniform performance with respect to the aging of components can be attained; and that troublesome design problems with the source of e_z can be avoided. The disadvantage of the circuit of Fig. 3 is that the inactive zone of the relays has a large and detrimental effect on the boundaries of the dead zone, whereas the inactive zone of the relays in Fig. 2 has a negligible effect on the dead-zone boundaries. For this reason, the design shown in Fig. 2 is generally preferable to that shown in Fig. 3.

Design Considerations

COMPARATOR

In many industrial cases, it is desirable to keep both the a-c input signal and the d-c input signal ungrounded. While it is easy to isolate the a-c signal by means of a transformer, it is not always possible to isolate the d-c input signal.

If the detector must be operated ungrounded, or grounded at some unfavorable point, the type of mounting and wiring must be given careful thought to

avoid spurious signals resulting from stray capacitance to ground. Resistors $R1$, $R2$, $R4$, and $R5$ may in some cases be selected to minimize some of this trouble. $X1$ is not critical; it is only necessary to select a rectifier with sufficiently high inverse voltage rating. In the 2-channel detector of Fig. 2, the dead-zone bias supply e_z, which must be isolated, will give considerable trouble as a source of unwanted signal. Fig. 4 shows a method of introducing a manually controlled reference point into the detector and another resistor arrangement for obtaining e_{g1}.

AMPLIFIER

Three factors are worthy of consideration: insensitivity to overloads, gain, and bandwidth. Figs. 5 and 6 are intended to serve as an aid to amplifier design. Fig. 5 shows the minimum required gain of the amplifier as a function of V/E, where V is an arbitrarily fixed value of e_{dc} and E is the peak value of e_{ac}. Since this curve is derived from a Fourier analysis of e_{g1}, as shown in the Appendix, it is the theoretical minimum required gain. To this curve must be added extra gain to compensate for losses and limited bandwidth, and for a safety factor. Fig. 6 indicates the minimum required bandwidth of the amplifier, and is a straightforward harmonic analysis of e_{g1} for several values of V/E.

RECTIFIER

Transient response of the detector may be dependent on the rectifier-filter circuit unless care is taken to insure a reasonably fast transient characteristic for the filter.

OUTPUT CIRCUIT

The relay circuit should be designed for fast transient response. The preferred type of relay is the type that operates a microswitch. The inexpensive sensitive relays will also work. In the event that the sensitive relay operates a larger contactor, this contactor may have to be of rugged design to withstand the severe operating conditions that may be imposed by the frequent corrective action of the detector.

REQUIRED   AMPLIFIER   GAIN

CALCULATED   AS   MINIMUM
TO   PRODUCE   1   VOLT   DC   AT e_{g2}
FROM   E = 1   VOLT   PEAK

ASSUMPTIONS
INFINITE   BANDWIDTH
PERFECT   RECTIFICATION

Fig. 5.   Required amplifier gain

APPLICATION

Fig. 7 shows the basic diagram of an existing application of the peak voltage detector to a variable-frequency alternator system. The detector automatically regulates the ratio of voltage to frequency to a constant preset value. The

Fig. 6.   Harmonic analysis of a rectified portion of a cosine wave

alternator has a power output rating of 300 kw, a frequency range of from 10 to 400 cycles per second, and a voltage range of from 0 to 600 volts rms.

Rectifiers $X11$ and $X12$ are type-6×5 tubes; this tube type was picked because its inverse voltage rating is sufficiently high for the application. The dead-zone bias supply e_z gave some design trouble because of undesired pickup and ripple; the entire system is ungrounded.   Electrostatically   shielded power transformers were found to be necessary to eliminate spurious signals.

The amplifiers consist of a half-section of 6SN7 operated at 105 volts plate voltage to prevent damage due to overloading. Each amplifier is transformer coupled to a resistance-capacitance rectifying network through a half-section of a 6H6. The rectified signals are applied to half-sections of a 6SN7, which operate the plate circuit relays CH1 and CH2.

System regulation of the voltage to frequency (E/F) ratio can be described in terms of e_z and e_{do}. Fig. 7 shows that the regulating system controls the output voltage (proportional to e_{ac}) but not the output frequency (proportional to e_{dc}). Output frequency is controlled by other means, external to the E/F control. Thus, if frequency is changed, either directly by the operator or indirectly as a result of increased output loading, the E/F control will change the value of output voltage to maintain the preset ratio of E/F within the voltage regulation limits, which are the limits imposed on e_{ac}. As shown in Fig. 2, the excursions of e_z, are limited to a maximum value equal to $(e_{do}+e_z)$ and to a minimum value equal to e_{do}, regardless of loading. System regulation can therefore be expressed as equal to or less than $e_z/(e_{do}+e_z)$. In the system of Fig. 7, e_z is preset to a constant value, while e_{do}, is controllable over a range of 40 to 1 for any given setting of the E/F ratio, and over a total range of 80 to 1 for all possible settings of the E/F ratio. This results in poor regulation at low frequencies, excellent regulation at high frequencies, and somewhat better regulation obtainable with high ratios of E/F than with low ratios. This type of regulation is well suited to the type of loading encountered in this application. For further information on variable-frequency alternator problems, see reference 3.

in the range $-\cos^{-1}(V/E)<\omega t<+\cos^{-1}(V/E)$, and as

$$f(\omega t)=0 \qquad (1)$$

elsewhere in the cycle.

The Fourier analysis states that

$$f(\omega t)=A_0+A_1 \sin \omega t+A_2 \sin 2\omega t+\ldots+ B_1 \cos \omega t+B_2 \cos 2\omega t+\ldots \quad (2)$$

where the amplitudes of the various harmonic components are

$$A_0=\frac{1}{2\pi}\int_{-\pi}^{+\pi}f(\omega t)d(\omega t) \qquad (3)$$

$$A_1=\frac{1}{\pi}\int_{-\pi}^{+\pi}f(\omega t)\sin \omega t d(\omega t) \qquad (4)$$

$$A_n=\frac{1}{\pi}\int_{-\pi}^{+\pi}f(\omega t)\sin n\omega t d(\omega t) \qquad (5)$$

$$B_1=\frac{1}{\pi}\int_{-\pi}^{+\pi}f(\omega t)\cos \omega t d(\omega t) \qquad (6)$$

$$B_n=\frac{1}{\pi}\int_{-\pi}^{+\pi}f(\omega t)\cos n\omega t d(\omega t) \qquad (7)$$

From consideration of symmetry, the right-hand members of equations 4 and 5 equal zero. Since $f(\omega t)=0$ except between $-\cos^{-1}(V/E)<\omega t<+\cos^{-1}(V/E)$, the lower limit of the integrals of equations 3, 6, and 7 becomes $-\cos^{-1}(V/E)$; the upper limit becomes $+\cos^{-1}(V/E)$. Substituting these limits and integrating yields the equations for the harmonic amplitudes as

$$A_0=\frac{E}{\pi}\sin\left(\cos^{-1}\frac{V}{E}\right)-\frac{V}{\pi}\cos^{-1}\frac{V}{E} \qquad (8)$$

$$B_1=\frac{1}{\pi}\left[E\cos^{-1}\frac{V}{E}+\frac{E}{2}\sin\left(2\cos^{-1}\frac{V}{E}\right)-2V\sin\left(\cos^{-1}\frac{V}{E}\right)\right] \qquad (9)$$

$$B_n=\frac{1}{\pi}\left[\frac{E}{(n-1)}\sin\left((n-1)\cos^{-1}\frac{V}{E}\right)+\frac{E}{(n+1)}\sin\left((n+1)\cos^{-1}\frac{V}{E}\right)-2\frac{V}{n}\sin\left(n\cos^{-1}\frac{V}{E}\right)\right] \qquad (10)$$

OTHER APPLICATIONS

The 4-channel peak voltage detector should be adaptable to the controller portion of a positional-type servo system that employs a 3-phase squirrel-cage induction motor as the power source of the controlled system. The slow speed zones of the detector in this system could be used to initiate plugging or braking of the motor shaft. The lower cost and greater reliability of the a-c squirrel-cage motor, when compared with the d-c motor that is more often employed in servos, makes this system particularly interesting. Methods for the design of such a system are given in references 4 and 5. The peak voltage detector is also applicable to heat controls when used in conjunction with resistance-type temperature detectors and to computer systems, since the multichannel detector may be considered to be an analogue-to-digital converter.

Appendix. Fourier Analysis of a Portion of a Cosine Wave

Fig. 6 shows that the function $f(\omega t)$ to be analyzed may be expressed as

$$f(\omega t)=E \cos \omega t-V \qquad (1)$$

Table I. Magnitudes of Harmonics

V/E=0.0		V/E=0.5		V/E=0.866		V/E=0.95		V/E=0.99		V/E=0.999	
n	$B_n/(E-V)$	n	$B_n/(E-V)$	n	$B_n/(E-V)$	n	$B_n/(E-V)$	n	$B_n/(E-V)$	n	$B_n/(E-V)$
0	+0.3183099	0	+0.2179955	0	+0.11060	0	+0.067275	0	+0.030025	0	+0.00949
1	+0.5000000	1	+0.391002	1	+0.22119	1	+0.13320	1	+0.055932	1	+0.01898
2	+0.2122066	2	+0.2757244	2	+0.19732	2	+0.12925	2	+0.059564	2	+0.01896
3	0	3	+0.1378322	3	+0.17147	3	0.122762	3	+0.058969	3	0.01895
4	-0.0424413	4	+0.0275665	4	0.13859	4	0.114086	4	0.058146	4	0.01892
5	0	5	-0.0275665	5	0.102879	5	0.103401	5	+0.057091	5	0.01889
6	+0.01818914	6	-0.0315045	6	0.067883	6	0.091798	10	0.048848	10	0.01861
7	0	7	-0.00984516	7	0.036743	7	0.078679	15	0.037000	20	0.01751
8	-0.01010508	8	+0.00984516	8	+0.011785	8	0.065712	20	0.0238442	30	0.01578
9	0	9	+0.01378322	9	-0.005715	9	0.052412	25	0.0117123	40	0.01356
10	+0.00643050	10	+0.00501208	10	-0.015993	10	0.039720	30	+0.000240	50	0.01104
		11	-0.00501208	11	-0.018705	11	0.027949	35	-0.003182	60	0.008395
		12	-0.00771089	12	-0.016615	12	0.017434	40	-0.0051389	70	0.005830
				13	-0.011305	13	+0.0084299			80	+0.003506
				14	-0.004787	14	+0.001096			90	+0.001555
				15	+0.001225	15	-0.0045003			100	+0.000058
				16	+0.005532	16	-0.008391			110	-0.000952
				17	+0.007565	17	-0.010642			120	-0.001496
				18	+0.0073557	18	-0.011308			130	-0.001634
				19	+0.005415	19	-0.011138				

Fig. 4, the required amplifier gain, is made by evaluating A_0 in equation 8 and then inverting A_0. Fig. 6 is obtained by evaluating equations 9 and 10 for the various harmonic amplitudes. Magnitudes of harmonics are shown in Table I.

References

1. SERVOMECHANISMS AND REGULATING SYSTEM DESIGN (book), H. Chestnut, R. W. Mayer. John Wiley and Sons, Inc., New York, N. Y., vol. 1, chap. 4, 1951.

2. PRINCIPLES OF SERVOMECHANISMS (book), G. S. Brown, D. P. Campbell. Ibid., chap. 7, 1948.

3. ADJUSTABLE FREQUENCY CONTROL OF HIGH-SPEED INDUCTION MOTORS, G. W. Heumann. AIEE Transactions, vol. 66, 1947, pp. 719–25.

4. A FREQUENCY RESPONSE METHOD FOR ANALYZING AND SYNTHESIZING CONTACTOR SERVO-MECHANISMS, Ralph J. Kochenburger. AIEE Transactions, vol. 69, pt. I, 1950, pp. 270–84.

5. RELAY SERVOMECHANISMS, T. A. Rogers, W. C. Hurty. Transactions, The American Society of Mechanical Engineers, New York, N. Y., vol. 72, no. 8, Nov. 1950 pp. 1163–72.

—————◆—————

Discussion

C. C Thomas (General Electric Company, Schenectady, N. Y.): Mr. Allen has presented a good clear picture of one method of designing a simple inexpensive comparison circuit which can be used as a component in certain industrial servo systems. His circuit is unique in that most industrial servo systems compare like quantities such as direct voltage against direct voltage or alternating voltage versus alternating voltage. It is rare that alternating voltage is compared to direct voltage especially in the more exacting servo systems.

Although not exact in detail, a common application using a similar comparison has been used in the design of our tracer control. Here it is desired to indicate when the deflection of the tracer exceeds a certain value. The deflection is indicated by a displacement in the position of two selsyns. This displacement produces an a-c error voltage. This alternating voltage is compared to (added to or subtracted from, depending upon the design) a fixed d-c bias voltage. The difference voltage operates on the grid of a vacuum tube to pick up a relay to shut down the machine. Here a variable alternating voltage was compared to a fixed direct voltage.

Another common application of this principle is in the rheostatic position follow-up. In this application one rheostat position represents the reference and the second rheostat the feedback of a servo system. A direct voltage is across each rheostat so that the error in position is represented by a direct voltage with the polarity indicating the direction of the error. This d-c error voltage is compared to an a-c bias voltage on each of two vacuum tubes and above a certain value turns on one tube for an error in one direction and the other tube for the opposite direction. Here a direct voltage in essence is compared to an a-c bias voltage.

Neither of these cases illustrate the exact method shown by Mr. Allen, who measures only the difference voltage on the positive half-cycle, and applies this difference to a tube grid. It is not clear why this is done. In many cases in industrial servos we use magnetic amplifiers instead of vacuum tubes, and would have extreme difficulty with such a comparison circuit in obtaining proper operation.

While many circuits of the same approximate cost and complexity as Mr. Allen's are used in industrial applications, it should by no means be inferred that all industrial servo systems are so simple. As a matter of fact, most industrial servo systems must be considerably more expensive and complex to obtain the performance now demanded by industry on a majority of applications.

•

Lloyd W. Allen: I am grateful to Mr. Thomas for his contribution to the subject, both by virtue of his comments on the peak voltage control and the description of a similar control made by his company. I am familiar with a version of the control described by Mr. Thomas, and have found it reliable and equal in simplicity to the peak voltage control. Either type of control can be used with equal results in some applications; each has peculiar advantages in other applications.

Thus, Mr. Thomas notes that the type of control he describes is better adapted to magnetic-amplifier circuits than is the peak voltage control; on the other hand, the peak voltage control may be better adapted to situations where advantage may be taken of two distinguishing features of the peak voltage control, namely the freedom of the input grid of the amplifier from high level signals e_{ac} and e_{dc}, and the freedom of the input grid from the d-c input voltage e_{dc}, which allows exploitation of the inherent stability of an a-c amplifier. Again, the control Mr. Thomas describes should be better adapted to situations requiring equal times for operation of the "raise" and "lower" relays, since it picks up either relay in the same manner to operate, whereas the peak voltage control picks up the "lower" relay but drops out the "raise" relay under similar operating conditions.

While the peak voltage control as described operates only on the positive half-cycle, in half-wave fashion, it is also possible, and may be preferable, to operate in full-wave fashion by using a bridge or other full-wave rectifier as a source of the a-c input.

A Method for the Preliminary Synthesis of a Complex Multiloop Control System

D. J. POVEJSIL
ASSOCIATE MEMBER AIEE

A. M. FUCHS
ASSOCIATE MEMBER AIEE

COMPLEX multiloop control systems may be synthesized by several different methods. The choice of synthesis procedure depends upon the configuration of the multiloop system. Fig. 1 shows a complex multiloop control system which has been reduced to several closed-loop within closed-loop systems. The synthesis of such a system can be performed by a straightforward application of the frequency response method.

Assume, e.g., that the over-all performance is specified and the characteristics of one block of the innermost loop are set. The innermost loop, loop 3, can now be designed to give an output-input response which would provide loop 2 with an equivalent fixed box, see Fig. 2, compatible with making the output-input of loop 2 more desirable as seen by loop 1. This process is continued until the last and outer loop meets the specified over-all performance.

Fig. 3 is a complex multiloop control system which has crosstalk terms. It is also characterized by having several inputs. It is desired to optimize the response to all of the input signals. This system can be reduced to the form shown in Fig. 2 for the relationship between any one input and one output. The entire system can be characterized, therefore, by a group of systems equivalent to Fig. 2. In general, there will not be a logical step-by-step synthesis process that can be developed from this presentation of the control problem. The complexity of the blocks of the "reduced" system often will mask the effect of a particular gain setting on the over-all performance. Nor does this procedure reveal the necessary added feedback terms which will bring about the specified performance most expeditiously. In fact, there is a tendency to select feedback quantities that simplify the analysis, i.e., that reduce the complexity of the block diagrams rather than to investigate other potentially more rewarding possibilities.

It is the purpose of this paper to present a method of performing the initial synthesis of a system which can be characterized by the block diagram of Fig. 3. By initial synthesis is meant:

1. Start with the differential equations that represent the uncontrolled systems, i.e., system to be controlled, and a set of over-all performance specifications.

2. Postulate the necessary feedback quantities which will bring about the desired over-all performance.

3. Establish gain settings for the feedback quantities.

4. Postulate the necessary compensating networks and set their time constants.

Inferred in steps 2, 3, and 4 is a continuous monitoring of the stability of the system.

Statement of the Problem

In the general case, the synthesis problem is presented in the following form.

1. The system to be controlled is characterized by n integro-differential equations relating each of the controlled variables (output quantities) to the input forcing functions, external disturbances, and initial conditions. The equations take this form

$$A_{11}C_1 + A_{12}C_2 + \ldots + A_{1n}C_n + F_{11}R_1 + F_{12}R_2 + \ldots + F_{1n}R_n + \Sigma(I.C.)_1 + \Sigma U_1 = 0$$

$$A_{21}C_1 + A_{22}C_2 + \ldots + A_{2n}C_n + F_{21}R_1 + F_{22}R_2 + \ldots + F_{2n}R_n + \Sigma(I.C.)_2 + \Sigma U_2 = 0 \quad (1)$$

$$A_{n1}C_1 + A_{n2}C_2 + \ldots + A_{nn}C_n + F_{n1}R_1 + F_{n2}R_2 + \ldots + F_{nn}R_n + \Sigma(I.C.)_n + \Sigma U_n = 0$$

where

A_{ij}, F_{ij}: analytic functions of the complex variable s

C_j = Laplace transform of the controlled variable (output), neglecting the initial conditions

R_j = Laplace transform of the input forcing function, neglecting initial conditions. R_j is limited to inputs which can be modified by feedback, as shown in Fig. 4, to form an actuating signal

ΣU_j = summating of external disturbances in the jth equations. The quantities cannot be modified directly by feedback to form an error signal

$\Sigma(I.C.)_j$ = summation of initial conditions in jth equation

These equations indicate that the number of independent controlled variables $(C_1 \ldots C_n)$ must be equal to or greater than the number of independent command signals $(R_1 \ldots R_n)$.

2. A set of performance specifications exist giving the desired relationship between the controlled variables and the command signals. There may also be performance specifications giving the desired controlled-variable response to the disturbing functions. In most instances a key part of this specification is that the characteristic equation (i.e., expansion of the system determinant) of the controlled system shall have a certain form; i.e.

$$s^m + B_{m-1}s^{m-2} + B_{m-2}s^{m-2} + \ldots B_1 s + B_0 = 0 \quad (2)$$

Where the characteristic equation is not explicitly defined in the performance specification, it can be deduced from other parts of the specification.

Note that the formulation of the desired characteristic equation includes the stability (damping) and bandwidth of the final system.

3. The relationship between a particular controlled variable and one of the command signals may be completely specified as

$$\frac{C_j}{R_k'} = \frac{G_{jk}}{\Delta} \quad (3)$$
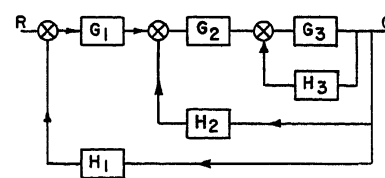
Fig. 1. Multiloop control system
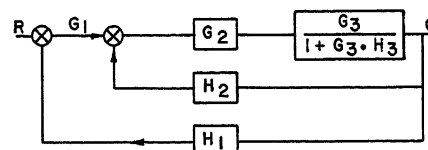
Fig. 2. First step in synthesis of the multiloop control system of Fig. 1
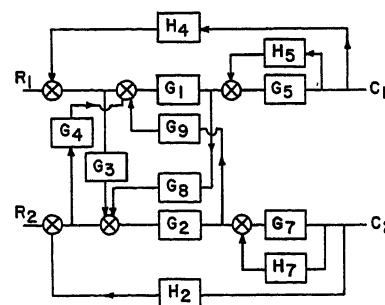
Fig. 3. Complex multiloop control system

This adds the problem of obtaining a particular $G_{jk}(s)$ to the synthesis problem. When the problem and the desired solution are expressed in this fashion, the synthesis of an adequate control system becomes possible in a logical, methodical manner as presented now.

Synthesis Procedure

The determinant of the uncontrolled system of equation 1 is

$$\Delta_u = \begin{vmatrix} A_{11} & A_{12}\ldots A_{1n} \\ A_{21} & A_{22}\ldots A_{2n} \\ \vdots \\ A_{n1} & A_{n2}\ldots A_{nn} \end{vmatrix} \quad (4)$$

where $\Delta_u =$ determinant of uncontrolled system. Expanding equation 4 into the characteristic equation of the uncontrolled system gives an equation of the form of equation 2. The coefficients of the uncontrolled characteristic equation will not match the specified coefficients. It is the purpose of the feedback and crossfeed terms to bring the controlled system's characteristic equation into agreement with the specified characteristic equation.

Referring to equation 1, it is seen that closing control loops around the uncontrolled system will have the effect of changing the form of the input forcing functions $(R_1 \ldots R_n)$ to command inputs $(R_1' \ldots R_n')$. The command input is related to the input forcing function by

$$\begin{aligned} R_1 &= R_1' + H_{11}C_1 + H_{12}C_2 \ldots + H_{1n}C_n \\ R_2 &= R_2' + H_{21}C_1 + H_{22}C_2 \ldots + H_{2n}C_n \\ &\vdots \\ R_n &= R_n' + H_{n1}C_1 + H_{n2}C_2 \ldots + H_{nn}C_n \end{aligned} \quad (5)$$

where

$R_n' =$ input to the controlled system before the summing point, as shown in Fig. 4
$H_{ij} =$ transfer function between measured controlled variable and the summing point; see Fig. 4

Now if the equations of 5 are substituted into equation 1, it is seen that the coefficients of the variables C_j assume the following form

$$\begin{aligned} A_{11}' &= A_{11} + F_{11}H_{11} + F_{12}H_{21} \ldots + F_{1n}H_{n1} \\ A_{12}' &= A_{12} + F_{11}H_{12} + F_{12}H_{22} \ldots + F_{1n}H_{n2} \quad (6) \\ &\vdots \\ A_{nn}' &= A_{nn} + F_{n1}H_{1n} + F_{n2}H_{2n} \ldots + F_{nn}H_{nn} \end{aligned}$$

Thus, the new determinant of the controlled system can be set up, and it can be expanded into the characteristic equation of the controlled system. The determinant of the controlled system will have the same general form as the determinant of the uncontrolled system. However, each of the self-coupling terms A_{jj}, and each of the mutual coupling terms A_{ij} has been altered by the addi-

tion of feedback terms. For example, if

$$A_{ij} = s^2 + as \quad (7)$$

the addition of the proper feedback terms might modify this parameter to the form

$$A_{ij} = s^2 + (a + a_1)s + b_1 + \frac{b_2}{s} \quad (8)$$

Note that the feedbacks have augmented an existing characteristic of the uncontrolled system by adding a_1. In addition, two new coupling terms, b_1 and b_2/s, have been created by the feedback.

Since it is possible to modify and change the self- and mutual couplings of the uncontrolled system in any desired manner, it is likewise possible to make the expansion of the characteristic determinant formed by these parameters fit the desired form of the specified characteristic equation given in equation 2.

In a similar manner, this synthesis procedure can be extended to the problem of meeting a particular output-input specification that might take the form of equation 3. In general, this will require controls in addition to those used to meet the basic specification on system response and damping. An example of such a control would be filtering of the reference input variable R_j'.

It is possible to build up a solution to the synthesis problem which evaluates each feedback term and it associated feedback transfer function in turn. After the gain and form of one feedback term has been established another one can be added and the procedure continued until the characteristic equation specification is met. The step-by-step synthesis procedure can be outlined as follows:

1. Compare the coefficients $B_{m-1}, B_{n-2}, \ldots B_0$ of the specified characteristic equation with the coefficients $B_{m-1}, B_{m-2}, \ldots B_0$ of the uncontrolled system.

2. Where the coefficients of the uncontrolled system are inadequate or nonexistent when compared to the desired control, examine the original system equations for the form of $F_{ij}H_{ji}$ which will establish the desired value of the coefficient. Careful tabulating procedures (of the factors of $B_{m-4}, B_{m-2}, \ldots B_0$) will permit a rapid selection of the A_{ij} term whose modification will produce the desired change in the determinant.

3. As each feedback term and its associated feedback transfer function are selected, it is necessary to re-evaluate the characteristic equation of the partially controlled system. This will reveal when the feedback quantity selected to improve one coefficient of the characteristic equation seriously interacts upon other coefficients. If the feedback quantity improves more than one coefficient of the partially controlled characteristic equation, it obviously represents a fortunate choice of feedback parameter; whereas, if the feedback term's beneficial effect is negated by its action upon another coefficient, it is rejected and other feedback terms are examined.

4. This process is continued until the coefficients of the controlled system's characteristic equation are equal to the coefficients of the specified characteristic equation within the allowable tolerances.

5. Where G_{jk} is also specified, as in equation 3, the process is repeated for the G_{jk} determinant in the same manner as for the characteristic determinant.

Several merits of the foregoing synthesis procedure are immediately evident:

1. The mutual coupling terms do not complicate the design procedure just described because these terms are manipulated in the same manner as the self-coupling (or main diagonal) terms. Thus, the designer is encouraged to examine the possibility of employing useful cross-coupling terms to achieve the desired control. This will enhance the possibility of achieving a minimum complexity control system, particularly when the designer possesses a good basic understanding of the uncontrolled system.

2. The stability of the over-all system is constantly being monitored as the control is being synthesized.

3. It is possible with this procedure to maintain a close relationship between the insertion of the feedback signals and their physical equivalent in the system being controlled. This results from the simplifications of the factors of the determinant and the method of relating them to the terms in the differential equation of the uncontrolled system.

4. This synthesis procedure accomplishes such a large part of the total synthesis that it greatly reduces the time that must be spent on an analogue computer to complete the synthesis. It rapidly reveals the unsatisfactory feedback terms which are not worth further investigation. It sets the gain and the form of the desired feedback terms. It leaves the analogue computer to study the effects of the nonlinearities and
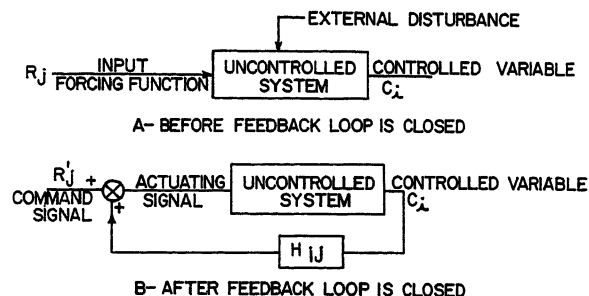
Fig. 4. Explanation of terminology

A- BEFORE FEEDBACK LOOP IS CLOSED

B- AFTER FEEDBACK LOOP IS CLOSED

the effects of the second-order delays in the feedback sensor and in the actuating error amplifier.

The entire synthesis procedure is frequently simplified by neglecting the delays present in the generation of the feedback quantities and in their amplification and application to the input forcing function. In systems where the actuator and sensor bandwidth are substantially faster than the system to be controlled (as a rate gyro and hydraulic actuator in an autopilot), this approximation does not cause any serious deterioration in the closeness of the results to the predicted values.

Addition of Shaping Networks

Frequently, it is not practical to achieve all of the desired performance without the use of shaping networks in the feedback transfer function. The addition of lead or integral networks raises the order of the characteristic equation. Where the specified characteristic equation is of higher order than the uncontrolled characteristic equation, this falls within the ordinary synthesis process as previously presented. The time constants of the shaping network can be selected as a separate step in arriving at the desirable coefficient of the characteristic equation.

When the addition of a compensating network is a compromise with generating a perfect derivative or integral function, the order of the controlled characteristic equation may exceed the order of the specified characteristic equation. This can be allowed for by adding an additional response term to the specified characteristic equation. An element of judgment concerning terms which can be neglected enters here.

Formulation of a Minimum Complexity System

One of the primary virtues of the synthesis procedure proposed in this paper is that it leads to the formulation of a set of analytical rules that define the general nature of a control system of minimum complexity in relation to its performance specifications. This can be shown in the following manner.

After the basic equations of the uncontrolled system are written the next step in the synthesis procedure is to establish analytical forms that satisfy the performance specification as shown by equations 2 and 3. Now, each specification can be considered separately. Observe that a specification of a characteristic equation of the form of equation 2 can

always be satisfied if the following requirements are met:

1. Since the desired characteristic equation has m variables, B_{m-1}, B_{m-2}, ... B_1, B_0, there must be at least m independent gains and time constants in the control systems.

2. The gains and time constants must be chosen such that each of the coefficients B_j can be expressed as a function of at least one of the independent gains or time constants of the control system.

3. Any combination of kB_j's must contain at least k number of the independent control system gains and time constants in such a form that k independent equations can be written.

Now, in some cases, it is possible that the constant of the uncontrolled system are such that less than m independent gains and time constants are required to meet the desired characteristic equation specification. Since the method of synthesis expresses the controlled system in the same terms as the uncontrolled, it is usually possible to recognize such a condition and to take advantage of it by the proper choice of controls.

Similar rules apply to specifications of the type given in equation 3. Here, a desired form is specified for the numerator of an output-input function in addition to the denominator requirement. Since the numerator G_{jk} can always be reduced to the form

$$G_{jk} = A_{m'}s^{m'} + A_{m'-1}s^{m'-1} + \ldots + A_1s + A_0$$

where $m' < m$, then by the same reasoning as before such a specification can always be met by adding $m'+1$ additional independent gains and time constants to the system. These quantities must meet the same three requirements with respect to the A_j coefficients as outlined for the B_j coefficients of the denominator.

Thus, the required complexity of the control system is directly related to the complexity of the original specification. When the controls are chosen in accordance with the foregoing rules, precisely the right number of gains and time constants can be postulated at the beginning of the synthesis, and redundant controls are eliminated.

Example: Lateral Aircraft Control System

To demonstrate this synthesis procedure, the design of an autopilot to control the lateral response of an aircraft is presented now. This synthesis problem is selected as an example because it does have large cross-coupling terms and it is frequently specified by the equivalent of a desirable characteristic equation. The lateral motion of the aircraft may be de-

scribed by three simultaneous linear equations, as follows (in the form of equation 1). See Fig. 5 for the physical interpretation of the variables in these equations.

Roll Moments

$$(s^2 - sl_p)\phi + (-s^2i_{xz} - sl_r)\psi + (-l_\beta)\beta + (-l\delta_a)\delta_a + (-l\delta_r)\delta_r = 0 \quad (9)$$

Yaw Moments

$$(-s^2i_{xz} - sn_p)\phi + (s^2 - sn_r)\psi + (-n_\beta)\beta + (-n\delta_a)\delta_a + (-n\delta_r)\delta_r = 0 \quad (10)$$

Side Forces

$$\left(-\frac{g}{V}\right)\phi + (1 - y_r)s\psi + (s - y_\beta)\beta + (-y\delta_a)\delta_a + (-y\delta_r)\delta_r = 0 \quad (11)$$

where

ϕ = roll angle, radians
ψ = yaw angle, radians
β = sideslip angle, radians
δ_a = aileron displacement, radians
δ_r = rudder, radians
V = aircraft forward velocity, feet per second
l, n, and y quantities are the aerodynamic coefficients
i_{xz} and i_{zz} terms are the product of inertia quantities

Expanding the determinant of the lateral motion gives a 5th order characteristic equation of the form

$$B_5s^5 + B_4s^4 + B_3s^3 + B_2s^2 + B_1s = 0 \quad (12)$$

where

$$B_5 = 1 - i_{xz}i_{zz} \quad (13)$$

$$B_4 = -l_p - n_r - y_\beta - n_pi_{xz} - l_ri_{xz} + i_{xz}i_{xz}y_\beta \quad (14)$$

$$B_3 = (1 - y_r)n_\beta + l_pn_r + l_\beta y_\beta + n_r y_\beta - n_pl_r + (1 - y_r)l_\beta i_{xz} + n_p i_{xz}y_\beta + l_r i_{xz}y_\beta \quad (15)$$

$$B_2 = -(1 - y_r)n_\beta l_p + (1 - y_r)l_\beta n_p - l_pn_ry_\beta + n_pl_ry_\beta - (g/V)l_\beta - (g/V)n_\beta i_{xz} \quad (16)$$

$$B_1 = (g/V)l_\beta n_r - (g/V)n_\beta l_r \quad (17)$$

The parameters used in this design are not equivalent to any particular airplane nor do they necessarily represent a realizable aircraft design. They have been selected to illustrate this synthesis procedure in a realistic fashion without introducing security restrictions upon the publication of the paper.

$l_p = -1.00$	$n_\beta = 6.41$
$l_r = 0.50$	$n\delta_a = -0.4$
$l\delta_a = -10.0$	$n_r = -4.0$
$l\delta_r = 0.10$	$g/V = 0.04$
$l_\beta = -10.0$	$y\delta_a = 0$
$i_{xx} = 0.01$	$y\delta_r = 0.03$
$i_{xz} = 0.05$	$y_\beta = -0.10$
$n_p = 0.10$	$y_r = 0.003$
$n_r = -0.25$	

This gives an uncontrolled characteristic equation of
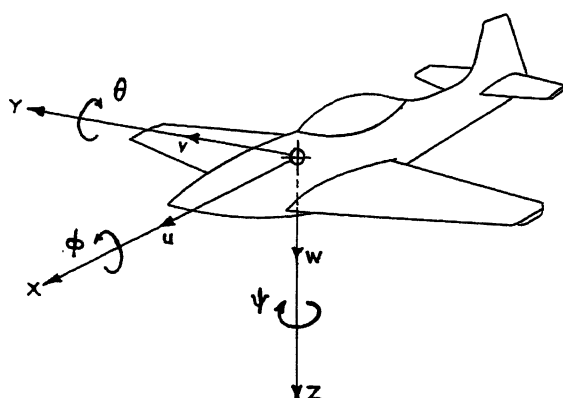
POSITIVE DIRECTIONS ARE SHOWN BY ARROWS

AXIS		FORCE	MOMENT ABOUT AXIS		ANGLE		VELOCITIES	
DESIGNATION		ALONG AXIS	DESIGNATION	POSITIVE DIRECTION	DESIGNATION	DEFINITION	LINEAR ALONG AXIS	ANGULAR ABOUT AXIS
X LONGITUDINAL		X	L ROLLING	Y→Z	ϕ ROLL	$\int p\,dt$	U	P
Y LATERAL		Y	M PITCHING	Z→X	θ PITCH	$\int q\,dt$	v	q
Z NORMAL		Z	X YAWING	X→Y	ψ YAW	$\int r\,dt$	w	r

Fig. 5. System of axes and symbols

X axis, in plane of symmetry and perpendicular to Z axis
Y axis, perpendicular to the plane of symmetry
Z axis, in plane of symmetry and perpendicular to the relative wind

Fig. 6. Lateral control system block diagram. Roll rate feedback to aileron; yaw rate feedback to rudder; yaw rate crossfeed to aileron

$$s^5+1.35s^4+6.61s^3+7.8s^2-0.028s=0 \quad (18)$$

The desired characteristic equation of the lateral mode of the aircraft is specified as

$$s(s^2+2\zeta_1\omega_{n1}s+\omega_{n1}{}^2)(s^2+2\zeta_2\omega_{n2}s+\omega_{n2}{}^2) \quad (19)$$

where

$\zeta_1=0.5 \qquad \omega_{n1}=2.5$ radians per second
$\zeta_2=1.0 \qquad \omega_{n2}=1.0$ radians per second

It is not suggested that this is a definitive statement of a desirable lateral autopilot response. It is offered as representative of a desirable response and as convenient to illustrate the design procedure. Substituting into equation 19, the values of the parameters just given results in desired characteristic equation of

$$s^5+4.5s^4+12.25s^3+14.75s^2+6.25s=0 \quad (20)$$

Now, it will be noticed that the desired characteristic equation 20 has the same order as the characteristic equation 18 of the uncontrolled system. This fact suggests that all of the required control may be obtained by modification of existing characteristics of the uncontrolled system. Furthermore, since the desired characteristic equation possesses four independent coefficients, four independent parameters are required in the control. An inspection of equations 13 through 17 shows that one combination of four parameters that will yield the required control is n_r, l_p, n_β, and l_a. Thus, the proposed control equations are

$$\delta_a=H_{11}s\phi+H_{12}s\psi \quad (21)$$

$$\delta_r=H_{22}s\psi+H_{23}\beta \quad (22)$$

When these equations are substituted into the original airplane equations, the modified aircraft parameters become

$$l_p=-(1+10H_{11}) \quad (23)$$

$$l_r=-(-0.5+10H_{12}-0.10H_{22}) \quad (24)$$

$$l_\beta=-(10-0.10H_{23}) \quad (25)$$

$$n_p=(0.10-0.40H_{11}) \quad (26)$$

$$n_r=-(0.25+4H_{22}+0.4H_{12}) \quad (27)$$

$$n_\beta=6.41-4H_{23} \quad (28)$$

$$y_r=0.003-0.03H_{22} \quad (29)$$

$$y_\beta=-(0.10-0.03H_{23}) \quad (30)$$

An exact solution for the values of H_{11}, H_{12}, H_{22}, and H_{23} can be obtained by substituting equations 23 through 30 into equations 13 through 17 with B_1 through B_5 defined by equation 20. For very complex problems, a digital computer is usually applied. However, in many practical cases, a systematic manual calculation procedure is possible by closing the loops in sequence and employing an iteration process suited to the particular system under consideration. For example, the following process can be used in this case:

1. For the first iteration assume that only l_p, l_r, n_r, and n_β are modified by the feedback terms. Using equations 13 through 17, the proper values of l_p, l_r, n_r, and n_β needed to fit B_1 through B_5 to the coefficients of equation 20 can be calculated by closing the H_{12}, H_{22}, H_{23}, and H_{12} loops in sequence.

2. Then, using these values of l_p, l_r, n_r, and n_β, the required control parameters H_{ij} can be calculated from equations 23, 24, 27, and 28.

3. New values of the parameters assumed to be constant in the first iteration, l_β, n_p, y_r and y_β, can then be calculated from equations 25, 26, 29, and 30.

4. Using the parameters obtained in step 3 as constants in equations 13 through 17, new values of l_p, l_r, n_r, and n_β can be found as the first step in the second iteration.

5. Repeat steps 2, 3, and 4 for as many iterations as are required to achieve the desired accuracy.

When this process was carried out for the example, the results of two iterations are as follows

$$H_{11}=0.125 \quad (31)$$

$$H_{12}=2.21 \quad (32)$$

$$H_{22}=0.2015 \quad (33)$$

$$H_{23}=0 \quad (34)$$

giving a controlled determinant of

$$s^5+4.50s^4+12.15s^3+14.73s^2+6.31s=0 \quad (35)$$

Note that, for the particular case chosen, one of the control parameters is zero. This implies that one of the existing parameters of the uncontrolled system needed no modification to achieve the desired control. Also observe that one of the key control terms was the crossfeed term H_{12}.

Fig. 6 is a block diagram of the resulting system. It reveals the startling complexity of this control as seen by the frequency response method.

Summary and Conclusions

A method has been presented for performing the initial synthesis of a multi-

loop, multiinput, and multioutput control system. The synthesis procedure starts by comparing the characteristic equation of the uncontrolled system with the specified characteristic equation. The minimum number of feedback terms and compensating networks necessary to match the two characteristic equations are found. An examination of the expanded form of the uncontrolled determinant reveals the fruitful feedback and compensating network terms. These are added to the system in such a way as to match the controlled system characteristic equation to the specified characteristic equation in a logical fashion.
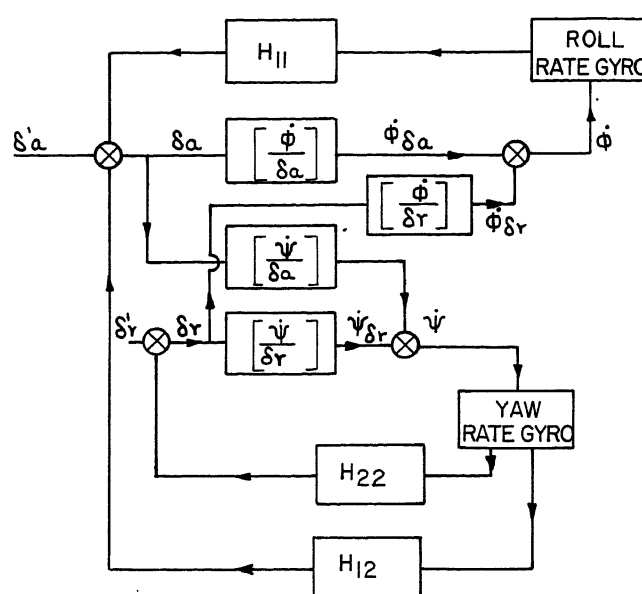
This method appears particularly applicable to autopilot, process control systems, engine controls, and fire control systems while it is less applicable to instrument servomechanisms and other single loop systems. However, several important steps must be taken to extend the utility of this method:

1. A more detailed study of the relationship between the form of the characteristic determinant and desirable transient system responses is required. There have been several papers analyzing this problem.[1] Generally the emphasis has been upon arriving at a damping value to minimize $\int_0^\infty f[e(t),t]dt$. None of these methods reveal the relationship between the natural frequencies of the characteristic determinant and the error transients. It is the basic problem of adequate system bandwidth

which plagues most system designers. Until this problem has more adequate treatment, the specification of the desired determinant will at best be a matter of good judgment rather than an exact procedure.

Note, however, that the need for a detailed understanding of the desirable form of the system determinant is inherent in every type of control system synthesis procedure. A frequency response specification, e.g., is a modified form of a characteristic equation specification. It is the determinant which bridges the gap between the transient behavior of a system and its frequency response. One of the strong points of this synthesis procedure is that it maintains a close relationship to the fundamental design objective, namely, that of a desirable transient behavior.

2. A careful study should be made to show how accurately a specified characteristic equation should be matched by the controlled system. This would assist in evaluating where the point of diminishing returns had been reached in adding additional feedback terms.

3. It may be possible to extend the describing function method of handling nonlinearities to this procedure. In either case, an analytic solution which includes the effect of nonlinearities would greatly assist in improving the validity of this design procedure.

It is interesting to note where the synthesis method presented in this paper differs from the classical transient design method.[2] In the classical approach the emphasis was upon maintaining a stable system as terms were added to the control system. In the method presented here, the stability is assumed once the desired charac-

teristic equation has been matched. The emphasis is now placed upon selecting a desirable response and upon synthesizing the minimum complexity control system rather than upon the stability of the system.

It is clearly understood by the authors that what has been presented here is not entirely new. These are overtones of Mr. Evans' root locus method and several other investigators have suggested a similar approach.[3,4] As far as the authors can find, however, this is the first time the emphasis has been placed upon a logical procedure which matches the characteristic equation term by term as related to the original set of differential-integral equations describing the uncontrolled system. It is also the first time this procedure has been suggested as end in itself worth pursuing and developing into a unique and powerful synthesis tool.

References

1. THE SYNTHESIS OF "OPTIMUM" TRANSIENT RESPONSE: CRITERIA AND STANDARD FORMS, Dunstan Graham, R. C. Lathrop. *AIEE Transactions*, vol. 72, pt. II, Nov. 1953, pp. 273–88.

2. BEHAVIOR AND DESIGN OF SERVOMECHANISMS, G. S. Brown. National Research and Development Committee, Washington, D. C., Nov. 1940.

3. CONTROL SYSTEM SYNTHESIS BY ROOT LOCUS METHOD, W. R. Evans. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 66–69.

4. AUTOMATIC FEEDBACK CONTROL (book), W. R. Ahrendt, John F. Taplin. McGraw-Hill Book Company, Inc., New York, N. Y., 1951.

Discussion

Dunstan Graham and Richard C. Lathrop (Wright Air Development Center, Wright-Patterson Air Force Base, Ohio): A majority of the literature on the analysis and synthesis of linear control systems deals extensively with the problem of obtaining the closed-loop response from a given open-loop response. Particularly the graphical methods associated with the names of Nyquist, Nichols, and Evans are employed every day by engineers for this purpose. The authors, however, have quite correctly pointed out the difficulties of attempting to synthesize multiloop, multichannel, coupled systems by methods which depend on successively closing loops with equivalent unit feedbacks. They deserve considerable credit for developing the mathematical relationship between the open- and closed-loop determinants of coefficients so as to show systematically the effects on the characteristic equation of the various possible feedbacks and cross-couplings for the more complex systems.

The claim, however, that "as far as the authors can find. . .this is the first time the emphasis has been placed upon a logical procedure which matches the characteristic equation term by term. . .etc." is likely to be misconstrued. In 1940 Imlay[1] examined the problem of lateral control of an aircraft

(the one used as an example by the authors), and solved for the feedback and cross-coupling terms necessary to match a desirable characteristic function term by term. In their 1944 monograph (published in English in 1948) Oldenbourg and Sartorius[2] suggest this method and illustrate its application to a third-order process control. In 1946 Whiteley[3] presented the method of "standard forms" (desirable characteristic functions) and illustrated its application to the synthesis of several different feedback controls. Whiteley's procedure has already been "suggested as an end in itself worth pursuing and developing" by, among others, Porter[4] and Jones and Briggs.[5]

With regard to the steps noted by the authors as required to extend the utility of the method:

1. The use of integral criteria has not by any means been restricted to the determination of damping factors. In reference 1 of the paper the characteristic functions were normalized for convenience, but, given sufficient adjustable feedbacks (assumed by the authors), a system of any desired bandwidth could be synthesized to conform to the characteristic function selected by some integral criterion. In the case where an insufficient number of parameters are adjustable, integral criteria may still be used to select optimum responses including system frequencies.[6] Oldenbourg and Sar-

torius[7] have suggested a new method for the determination of desirable transfer function coefficients including the frequency-determining ones. The coefficients are developed from the "first principle" of flat frequency response and are similar in many cases to the ones selected by integral criteria.

2. Reference 1 of the paper has presented a study of how accurately a specified characteristic equation should be matched by the control system. In general, the coefficients of the low powers of the complex variable must be matched as closely as the vagaries of physical equipment (including high-quality analogue computers) permit; while the system dynamic characteristics are somewhat indifferent to changes in the coefficients of the high powers of the complex variable. These statements are especially true of higher-order systems.

3. It would be a major advance in control system synthesis if the method described by the authors could be extended to cover nonlinearities. Have they had any success with this since the paper was written?

In our opinion the method which the authors have contributed is a powerful one, and it will see increasing application to the design of complex systems, such as aircraft automatic controls. It is now feasible to synthesize coupled loops simultaneously. This cannot be done with frequency-response or root-locus techniques.

REFERENCES

1. A Theoretical Study of Lateral Stability with an Automatic Pilot, F. H. Imlay. *National Advisory Committee for Aeronautics TR 693*, Washington, D. C., 1940.

2. The Dynamics of Automatic Controls (book), R. C. Oldenbourg, H. Sartorius. The American Society of Mechanical Engineers, New York, N. Y., 1948.

3. The Theory of Servo Systems, with Particular Reference to Stabilization, A. L. Whiteley. *Journal*, Institution of Electrical Engineers, London, England, vol. 93, 1946, pp. 353–72.

4. Introduction to Servomechanisms (book), A. Porter. Methuen and Company, Ltd., London, England, 1950.

5. A Survey of Stability Analysis Techniques for Automatically Controlled Aircraft, A. L. Jones, B. R. Briggs. *National Advisory Committee for Aeronautics TN 2275*, Ames Aeronautical Laboratory, Moffett Field, Calif., Jan. 1951.

6. The Influence of Time Scale and Gain on Criteria for Servomechanism Performance, Dunstan Graham, Richard C. Lathrop. *AIEE*
Transactions, vol. 73, pt. II, July 1954, pp. 153–58.

7. A Uniform Approach to the Optimum Adjustment of Control Loops, R. C. Oldenbourg, H. Sartorius. *Transactions*, The American Society of Mechanical Engineers, vol. 76, no. 8, Nov. 1954.

D. J. Povejsil and A. M. Fuchs: Mr. Graham and Mr. Lathrop have brought up a number of important points. They are quite correct in stating that our statements concerning the originality of our approach are misleading. The references cited as previously employing the essential elements of this design procedure add to our own knowledge of the history of this approach.

With regard to the problem of bandwidth it is true that any desired bandwidth may be synthesized if a sufficient number of parameters are adjustable. However, the problem is to determine what bandwidth is desired in a given application—a problem that will generally require considerable engineering judgment.

The discussers' point concerning the insensitivity of the system dynamic characteristics to changes in the coefficients of the high powers of the complex variable is well taken. This fact makes it possible to simplify the preliminary synthesis procedure by neglecting high-frequency terms, a procedure that was actually used in the example in the paper. However, such a simplification is not required by the method. One advantage of the method is that the desired characteristics of the high-frequency modes may be synthesized in the same manner used for the low-frequency modes. This fact is particularly useful for the synthesis of automatic flight control systems. The increasing natural frequencies of fighter aircraft have made it difficult to justify the convenient fiction that control surface servo drive systems are appreciably more rapid than the aircraft they are controlling.

We have attempted to extend the method to nonlinear systems. At the present time no promising approach has been found.

A General Theory for Determination of the Stability of Linear Lumped-Parameter Multiple-Loop Servomechanisms (and Other Feedback Systems)

THOMAS S. AMLIE
NONMEMBER AIEE

THOMAS J. HIGGINS
MEMBER AIEE

THE main purpose of this paper is to set out the details of a general theory enabling direct determination of the stability of linear lumped-parameter multiple-loop servomechanisms (and other feedback systems) which is broader in scope, more direct in application, more powerful in use, and yields greater insight to the physical functioning of both the system as an entity and its subloops than does presently available theory.

The practical importance of the content of this paper is evidenced in Bothwell's[1] recent statement to the effect that satisfactory determination of the stability aspects of a servo system comprises the prime problem of design. Obviously, if the system is to perform effectively in practice, it is essential that the system be stable as an entity. In multiple-loop servo systems, this over-all requirement necessitates, in turn, implicit or explicit knowledge of the effects on over-all stability which result from a change of structure of the linear loops. Additively, and often, explicit knowledge of the nature

of the stability of one or more inner loops is needed, e.g., through the requirement that loops which incorporate devices which might be damaged by loop overloading due to loop instability resulting from failure (in whole or in part) of components within the loop must be safeguarded against such possibility. In consequence, for these and other reasons, a body of theory ought to be available which enables ready, inclusive analysis of the stability of multiple-loop servo systems on either an over-all or a per-loop basis. Although a certain amount of general theory which can be utilized for such purpose is scattered through the literature of classical dynamics and feedback amplifier theory, no critically considered, integrated account of application of it to the analysis of multiple-loop servo systems is to be found. Moreover, the various theories available in other sources prove, when gathered into an integrated whole, insufficient to enable analysis of limitedly stable systems. Yet inasmuch as the consideration of limited stability

(defined in Section 1) is a necessity in certain types of guided missile controls, autopilots, and other aero-device control systems, it is obviously desirable that this gap be filled by an extension of existing theory. The practical importance of this paper thus stems from the fact that its contents comprise both an integrated account of conventional multiple-loop servo system theory and the desired extension of it which enables analysis of limitedly stable systems.

The subject falls naturally into three rather distinct parts:

1. A statement of the general definition of stability, with definitions of the three major subclasses of stable, unstable, and limitedly stable systems.

2. An integrated account of general stability analysis of single-loop servo systems.

3. An extension of this general single-loop theory to the analysis of multiple-loop servo systems on an over-all basis.

The theory of the first portion of the paper is directly applicable to determina-

tion of the stability of single-loop systems; and as extended to multiple-loop systems yields knowledge of the stability on an over-all basis between specified input and output points. A theory which enables direct investigation of a multiple-loop system on a per-loop basis can be founded on a generalization of Bode's[2] method of return differences as used in feedback amplifier theory. Accordingly, in the second part of the paper Bode's method of return differences is described with a generalization of it which enables complete determination of the stability of multiple-loop systems.

Bode's method of return differences, and its generalization, are based on mesh (or nodal) analysis as effected by the use of complex impedances (or admittances). This approach is convenient for feedback amplifier design, for which purpose Bode originated his procedure, since the descriptive equations of such systems are usually couched in terms of complex impedances (or admittances). Currently, however, servomechanism theory is largely given in terms of Laplace-transform transfer functions. Obviously, therefore, it is most desirable to have a generalized theory akin to that developed in the second portion of the paper, but in terms of transfer functions. Therefore, the third part of the paper comprises an original development of this theory. An application of this theory is illustrated by solution of a 3-loop servomechanism system.

Discussion is restricted to lumped-parameter systems because of space limitations. However, the application of the theory so advanced to systems encompassing distributed parameters is immediately possible by well-known theory.

1. Fundamental Concepts Underlying Stability Theory

BASIC DEFINITIONS

An important part of servomechanism theory is the determination of the stability of a specified servo system. Consider a linear system which is either in a quiescent state or, because of the action of certain impressed forces, in a constant (unvarying), a periodic, or a transient state. Let this system be subjected to additional disturbing forces which come on at a certain time $t = -t_1 (t_1 > 0)$; act for a finite length of time; and then vanish at, say, $t = 0$. The question is now posed: Will the system ultimately (i.e., as $t \to \infty$) achieve the state it would have achieved as $t \to \infty$ if the disturbing forces had not acted? If so, then it can be said that

"the system eventually returns to its original ultimate state."

To determine the answer to this question, the nature of the stability of the system must be ascertained. A system is characterized as:

1. Stable if it returns to its original ultimate state.

2. Unstable if the ultimate value of one or more of the dependent variables (responses) $c(t)$ differs from the corresponding original ultimate value by either an arbitrarily large value or by an oscillatory function of arbitrarily large maximum amplitude.

3. Limitedly stable if the ultimate value of one or more of the dependent variables (responses) $c(t)$ differs from the corresponding original ultimate value by a constant value, or a periodic function of finite maximum amplitude, or a superposition of the two.

Thus, consider a disturbed system of one dependent variable, where $c(t)$ denotes the disturbed response of this variable and $c_0(t)$ denotes the corresponding original (undisturbed) response. Then the system is:

1. Stable if $\lim_{t \to \infty} [c(t) - c_0(t)] = 0$.

2. Unstable if $\lim_{t \to \infty} [c(t) - c_0(t)] = \infty, -\infty$ or an oscillatory function of arbitrarily large maximum amplitude.

3. Limitedly stable:
   a. With constant displacement if $\lim_{t \to \infty} [c(t) - c_0(t)] = K_0$.
   b. With sustained oscillation if $\lim_{t \to \infty} [c(t) - c_0(t)] = P(t)$.
   c. With both if $\lim_{t \to \infty} [c(t) - c_0(t)] = K_0 + P(t)$.

Here K_0 designates a constant value and $P(t)$ a periodic function of finite maximum amplitude. The formulation of a similar classification for a system of two or more dependent variables is possible.

Further subdividing the concept of stability, a linear system is characterized as:

4. Absolutely stable with respect to a parameter of the system if it is stable for all values of this parameter (commonly, this characterization is made with respect to only positive values of the parameters).

5. Conditionally stable with respect to a parameter if the system is stable for only certain bounded ranges of values of this parameter (commonly, this characterization also is made with respect to only positive values of the parameter).

Finer distinctions of stability of a linear system are to be found in the literature. However, the five definitions given here are sufficiently discriminating to cover most of the differences.

In reading textbooks or periodicals the following points must be kept in mind with respect to these definitions.

First, most technical writers characterize a system as only stable or unstable, the concept of limited stability not being distinguished. Thus, a writer essentially interested in a system which is considered satisfactory in performance only if it "returns to its original state" will consider a limitedly stable system as unstable. On the other hand, if the system is considered satisfactory in performance provided disturbance does not cause it to "run away" (i.e., causes the value of one or more of the dependent variables to differ, within prescribed allowable limits, from the corresponding original ultimate value by a constant value or an oscillatory value of limited maximum amplitude as $t \to \infty$), the writer will characterize a limitedly stable system as stable. (Some writers say such a system "operates at the limit of stability.") The reader should bear in mind the first three of the five categories of stability discussed here, and interpret limited stability in accordance with the writer's use of stable or unstable.

Second, although most writers use the term "conditionally stable" in the sense defined in item 5, some use it to denote what is defined in item 3 as a limitedly stable system. In view of the varied usage of this term, the reader must ascertain from the context the writer's meaning of conditionally stable.

BASIC PROCEDURES FOR DETERMINING STABILITY

The analytical determination of the stability of a linear system can be simplified as follows. Consider, first a system characterized by one response, thus by one dependent variable $c(t)$. If the system is originally quiescent at $t = -t_1$ ($t_1 > 0$), the disturbing forces coming on at $t = -t_1$ act alone. If on the other hand, at $t = -t_1$ the system is originally in a constant, a periodic, or a transient state, the disturbing forces act in conjunction with the original forces producing this state. In that the system is linear, superposition of disturbing and original forces and corresponding responses can be utilized to calculate the resulting disturbed response $c(t)$.

Thus, the resulting response $c(t)$ can be calculated as the sum of the undisturbed (original) response $c_0(t)$ due to the original forces acting alone and the disturbing response $c_d(t)$ produced by the disturbing forces, considered as acting alone on the system taken as quiescent at the time the disturbing forces are impressed on the system. Hence, $c(t) = c_0(t) + c_d(t)$. The original response $c_0(t)$ is known (in that, if desired, it can be

calculated by well-known theory for any specified system and original forces). Accordingly, determination of the disturbed response $c(t)$ of the system reduces, in essence, to determination of the disturbing response $c_d(t)$.

Now the disturbing forces come on at $t=-t_1(t_1>0)$ and vanish at, say, $t=0$. At this instant $t=0$, both $c_d(t)=c_d(0+)$ and the values of its derivatives are known (in that they can be calculated by well-known theory for any specified system and disturbing forces). Accordingly, if these known values are considered as initial values of $c_d(t)$ at $t=0$, the disturbing response $c_d(t)$ subsequent to $t=0$ is precisely that which results from finding the response $c_{ff}(t)$ of the system considered as force-free but subject to the just-stated initial values of $c_d(t)$ at $t=0$.

Hence, subsequent to removal of the disturbing forces at $t=0$, the disturbed response of the system after $t=0$ can be considered as the sum of the known original response $c_0(t)$ and the force-free response $c_{ff}(t)$. Thus, $c(t)=c_0(t)+c_{ff}(t)$, for $t>0$. But by virtue of the analytic characterization of stability under items 1, 2, and 3 of Section 1, the nature of the stability is given by determination of the nature of $[c(t)-c_0(t)]-c_{ff}(t)$ as $t\to\infty$. Thus, the system is stable if $\lim\limits_{t\to\infty} c_{ff}(t)=0$, etc., as specified in items 1, 2 and 3. Accordingly it follows that determination of the stability of a linear system reduces to determination of the ultimate nature of the force-free response $c_{ff}(t)$; i.e., to determination of the ultimate response of the force-free system subject to a certain specified set of initial values $c_d(0+)$, $c_d'(0+)$,... at $t=0$. This calculation is a well-known problem in the field of transient analysis of linear systems, easily effected by use of the Laplace transform or other operational analysis.

A somewhat less general formulation, but a more convenient one in that it enables simpler calculation when applicable, is as follows: Let the system be of such nature that the ultimate value of the force-free state $c_{ff}(t)$ is independent of the initial conditions at $t=0$, and thus of the nature of the disturbing forces.* Consider the disturbing force $f_d(t)$ as impressed at $t=-t_1(t_1>0)$, vanishing at $t=0$, and specified over the range $-t_1\le t\le 0$ by a function of such nature that

$$\lim_{t_1\to 0}\int_{-t_1}^{0} f_d(t)dt=1$$

If now $t_1\to 0$, the disturbing force approximates in the limit to a unit impulse $\delta(t)$, acting at $t=0$ (its Laplace transform is $L[\delta(t)]=1$). But in such a case cal-

culation of the disturbing response $c_d(t)$ not only reduces to determination of the response of the quiescent system acted on by the impressed force $\delta(t)$ at $t=0$ but $c_d(t)\equiv c_{ff}(t)$. Correspondingly, determination of stability of the system reduces to determination of the ultimate nature of the response of the quiescent system to unit impulse $\delta(t)$. This calculation is also a well-known problem in the field of transient analysis of linear systems (specifically, the desired response is the so-called Green's function of the system).

This second statement of the stability problem is somewhat simpler than the first, for determination of initial conditions at $t=0$ prior to calculation of $c_{ff}(t)$ is obviated. However, it is correct only if the ultimate value of $c_d(t)$ is independent of the nature of the disturbing forces.*† This second formulation is the one advanced in most servomechanism literature, although the writers usually fail both to state the restriction just mentioned or even to delineate clearly what they mean by stability of a system. For example, various textbooks on servomechanism theory wrongly advance, as illustrative of their previous discussion of stability, a consideration of the response of a simple system to a step function. But such an example is illustrative of the ultimate response of a forced system (discussed in Section 4); whereas actually the concern of the authors at this point and throughout the textbook is with the stability of a disturbed system, per the second formulation advanced in the foregoing. As certain systems may be stable as far as disturbed response is concerned, but unstable in so far as forced ultimate response is concerned, it is obvious that entirely different considerations are concerned in the two cases, and each must be analyzed accordingly.

2. Analytic Determination of Stability of Lumped-Parameter Single-Loop Linear Systems

DETERMINATION OF DISTURBING RESPONSE $c_d(t)$

Consider, then, determination of the actual disturbing response $c_d(t)$ subsequent to $t=0$ as calculation of the response $c(t)=c_d(t)$ of a lumped-parameter linear system of one dependent variable

---

*Such independence is commonly, but not always, true of linear systems; in particular, it is true for all dissipative systems, but not for all nondissipative systems.

†However, as the former, rather than the latter type of system is commonly encountered in practice, this slightly less general formulation suffices for most stability analyses.

Table I. Contribution to $c_u(t)$ of Six Varieties of Roots

Nature of a,b	Kind of Root	Contribution to Ultimate State
$a<0$, $b=0$	positive real root, $m\ge 1$	∞ or $-\infty$
$a>0$, $b=0$	negative real root, $m\ge 1$	0
$a=0$, $b=0$	zero real root, $m=1$	K_0
$a=0$, $b=0$	zero real root, $m\ge 2$	$\infty^{m=1}$ or $-\infty^{m-1}$
$a=0$, $b\ne 0$	conjugate imaginary roots, $m=1$	$K_0\sin bt$
$a=0$, $b\ne 0$	conjugate imaginary roots, $m\ge 2$	oscillatory function of infinite maximum amplitude
$a>0$, $b\ne 0$	conjugate complex roots, real part positive, $m\ge 1$	oscillatory function of infinite maximum amplitude
$a<0$, $b\ne 0$	conjugate complex roots, real part negative, $m\ge 1$	0

$c(t)$, free of external forces, but at $t=0$ energized by a certain set of initial conditions; or, alternatively, consider determination of the response of this system which, originally quiescent, is energized at $t=0$ by a unit impulse $\delta(t)$. The Laplace transform $C(s)$ of $c(t)$ can in either case be expressed as a rational fraction in s: thus, as $C(s)=A(s)/B(s)$, where $A(s)$ and $B(s)$ are polynomials in s and have no factor, and thus no zeros, in common.

On expressing $C(s)=A(s)/B(s)$ as a sum of partial fractions and inverting to obtain $c(t)$ through use of well-known theory, it is found that:

1. Each distinct real root $s=a$ of the characteristic equation $B(s)=0$ contributes a term, $K_0\epsilon^{at}$.

2. Each real root $s=a$ of multiplicity m contributes $\sum\limits_{i=0}^{m-1} K_i t^i \epsilon^{at}$.

3. Each distinct pair of conjugate imaginary roots $s=\pm jb$ contributes $K_0\sin bt$.

4. Each pair of conjugate imaginary roots $s=\pm jb$ of multiplicity m contributes $\sum\limits_{i=0}^{m-1} K_i t^i \sin bt$.

5. Each distinct pair of complex roots $s=a\pm jb$ contributes $K_0\epsilon^{at}\sin(bt+\theta_0)$.

6. Each pair of complex roots $s=a\pm jb$ of multiplicity m contributes $\sum\limits_{i=0}^{m-1} K_i t^m \epsilon^{at}\sin(bt+\theta_i)$.

It follows from items 1 to 6 that in general $c(t)$ will comprise a sum of terms, each containing ϵ^{at} where a is negative, the limit of which sum approaches zero as $t\to\infty$, the sum of the remaining terms. The first sum is called the transient-state component $c_t(t)$ and the second sum the ultimate-state component $c_u(t)$. Hence, in general $c(t)=c_t(t)+c_u(t)$. The contribution to $c_u(t)$ of each of the six varieties

of roots listed (as found by letting $t \to \infty$) is given in Table I.

By compounding the contributions of each of these roots of $B(s) = 0$ as indicated in Table I, the nature of the ultimate state of the force-free system with prescribed initial values or of the quiescent system subjected to unit impulse $\delta(t)$ at $t = 0$ is determined, and therefrom the stability of the system. Thus, e.g., if all the roots of $B(s) = 0$ are real and negative, the disturbing ultimate response is zero; hence the ultimate state of the disturbed actual system (which is a superposition of the original state plus the disturbing state) approaches the original ultimate state of the system at $t \to \infty$: i.e., after the disturbing forces disappear, the system ultimately returns to its original state, hence is stable.

GENERAL STABILITY CRITERIA

By analyzing the various possibilities and by agreeing to term any root as complex, but distinguishing among complex roots as to whether they have a positive real part, a negative real part, or a zero real part, the following *root* criteria are obtained: The system is stable if all roots of the characteristic equation $B(s) = 0$ have negative real parts; is unstable if any root has a positive real part, or if any root has a zero real part and is of multiplicity greater than one; and is limitedly stable if one or more roots has a zero real part and is of multiplicity one, the remaining roots having negative real parts.

Interpreted in terms of the location of the roots in the complex plane, the following location criteria are: The system is stable if all roots of $B(s) = 0$ lie to the left of the imaginary axis; is unstable if any lie to the right of the imaginary axis or if any of multiplicity greater than one lie on the imaginary axis; and is limitedly stable if any of multiplicity one lie on the imaginary axis and the remaining roots lie to the left of this axis.

In the limitedly stable case, a root at the origin results in superposition of a constant value on the original ultimate state, whereas a pair of conjugate roots on the imaginary axis results in superposition of a periodic oscillation of finite maximum amplitude on the original ultimate state.

METHODS FOR DETERMINING NATURE OF ROOTS OF CHARACTERISTIC EQUATION

In examining a specified system one often is interested only in ascertaining whether or not the system is stable. To this end it is desirable to have a means of determining whether or not all of the roots of the characteristic equation $B(s) = 0$

are pseudonegative (i.e., have negative real parts) without the necessity of determining the actual numerical values of these roots, which procedure may require considerable numerical labor if the equation is of higher order than a quadratic. There are numerous methods for determining whether or not all of the roots of a polynomial equation are pseudonegative, which can be characterized, roughly, as either algebraic or nonalgebraic, depending on whether algebraic computation alone or a conjunction of algebraic computation and other calculation such as graphical delineation or some phase of complex variable theory is used in applying them.

The methods most commonly cited in textbooks on servomechanism theory are those of Hurwitz, Routh, and Nyquist. The closely related methods of Hurwitz and Routh are algebraic in nature, involving only routine numerical computation. Nyquist's is a nonalgebraic procedure, somewhat more complex to use.

3. Analytic Determination of Stability of Multiple-Loop Systems

To determine the stability of multiple-loop lumped-parameter systems only a slight extension of the foregoing remarks on single-loop systems is necessary. Instead of one, we have a system of differential equations characterizing the system. In general, if no constraints are imposed on the system the number of equations, and thus the number of independent loop variables or responses $c_j(t)$, equals the number of loops. Let us consider such here.

On transforming the set of differential equations stemming from either the first or second formulation of the stability problem and solving the resulting set of algebraic equations for a particular loop variable $C_j(s)$, $C_j(s)$ is obtained, expressed as $C_j(s) = D(s)/\Delta(s)$. Here $\Delta(s)$ is the determinant of the coefficients of the equations in terms of the $C_j(s)$, and $D(s)$ designates the determinant formed by replacing the column of coefficients of $C_j(s)$ by the column stemming from the initial conditions (first formulation) or the disturbing unit impulse (second formulation). After eliminating factors common to $D(s)$ and $\Delta(s)$ an expression of the form $A(s)/B(s)$ results and analysis proceeds as before where, of course, each of the $C_j(s)$ must be considered before statement of stability can be formulated.

A rather recondite point of multiple-loop stability theory should be mentioned here. It would be natural to surmise that the nature of the stability depends on the

roots of $\Delta(s) = 0$; in fact, however, it depends on the roots of $B(s) = 0$. Thus, it is not difficult to construct systems where $\Delta(s) = 0$ has one or more repeated roots; yet because of the fact that $D(s)$ has factors in common with $\Delta(s)$, in a certain manner, $B(s)$ has no repeated root; whence $c_j(t)$ contains no power of t, and thus does not contribute to absolute instability. It is easily established (as was first explicitly done by Routh[3,4]) that if $\Delta(s) = 0$ has a root $s = a$ of multiplicity m, then a necessary and sufficient condition that no power of t is to be contributed by this root is that all the first, second, etc., minors of $\Delta(s)$ up to and including the $(m-1)$th minor shall vanish for $s = a$. Accordingly, when investigating stability in a multiple-loop system it is necessary to make certain whether or not $D(s)$ and $\Delta(s)$ have factors in common, and if so, to proceed accordingly.

4. Ultimate State of Forced Systems

Although the concept of stability as given here is usually considered as a property of the system with respect to disturbance (i.e., one is to investigate whether or not the disturbing values produced by $r(t) = \delta(t)$ ultimately vanish and the system returns to its original ultimate state) much of the associated theory can be applied to determine the ultimate nature of a response to a forcing function; i.e., whether the ultimate response is infinite, zero, of finite constant value, an oscillatory function of finite or infinite amplitude, etc. Thus, in the specific case of a single-loop servo system of closed-loop transfer function $1/Z(s)$, the operational response to a reference input $r(t)$ is

$$C(s) = \frac{R(s)}{Z(s)} + \frac{\phi(s)}{Z(s)}$$

where $\phi(s)$ stems from the initial conditions. On eliminating (or combining) factors common to $R(s)$ and $Z(s)$ and to $\phi(s)$ and $Z(s)$, the operational response takes the form

$$C(s) = \frac{A(s)}{B(s)} + \frac{A'(s)}{B'(s)}$$

and the analysis outlined in Section 2 can be applied as pertinent to determine the ultimate nature of $c(t)$.

5. General Stability Analysis of a Linear Lumped-Parameter Multiple-Loop Active Network

GENERAL NETWORK EQUATIONS

The procedure outlined in Section 3 results in specification of stability on an
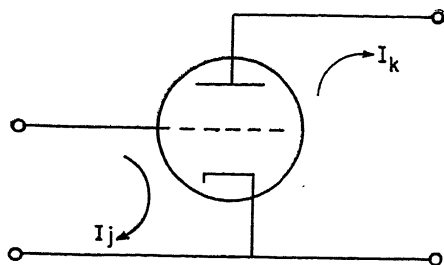
Fig. 1. Schematic of a triode

over-all basis. Again, it is easiest applied if the complete system is known analytically. However, this is not always feasible in practice; e.g., the transfer functions of certain components may not be known analytically. Accordingly, it is advantageous both for this and other reasons (such as stated in the second paragraph of the paper) to have available a theory for determining the stability of multiple-loop systems which is applicable to both 1. when complete analytic description is available, or when a system is known, in whole or in part, through physical realization rather than through analytic description; and 2. when knowledge of the stability of certain specified inner loops as well as over-all stability is desired. This theory will now be described.

It is easiest developed and grasped through analogy with the somewhat similar theory based on return-voltage differences which Bode developed for analysis of the conventional stability of multiple-loop feedback amplifiers. The only readily available, comprehensive account of this theory appears to be that advanced in Bode's book, and this account is rather loosely phrased and somewhat difficult to follow (at least, so it appears to the authors). Therefore a concise, but clear-cut, delineation of Bode's method of return differences is now given for determining the conventional stability of linear lumped-parameter multiple-loop (electric) systems as a necessary preliminary to a generalization of this theory to encompass analysis of limitedly stable systems and subsequent development of a similar theory for multiple-loop servomechanisms.

A quiescent n-loop electric network comprised of active and passive elements can be characterized either by a set of mesh or nodal operational equations. Obviously, that set might well be chosen which provides the simpler system of operational equations. Assume for the moment that the system is such that it is best characterized by its set of operational mesh equations; thus, by

$$
\begin{aligned}
Z_{11}I_1 + Z_{12}I_2 + \ldots + Z_{1n}I_n &= E_1 \\
Z_{21}I_1 + Z_{22}I_2 + \ldots + Z_{2n}I_n &= E_2 \\
&\vdots \\
Z_{n1}I_1 + Z_{n2}I_2 + \ldots + Z_{nn}I_n &= E_n
\end{aligned} \tag{1}
$$

where I_j, E_j, and Z_{ij} denote Laplace transforms, thus functions of the complex variable s.

In equation 1 assume that the Z_{ij} represent bilateral passive impedances; for such the reciprocity condition holds: i.e., $Z_{ij} = Z_{ji}$. Next, in equation 1 let all of the driving voltages E_i except E_1 be either zero or voltages produced in unilateral coupling impedances. Thus, typically, $E_k (k \neq 1)$ might represent the voltage produced in the kth mesh by a current I_j in the jth mesh; as, e.g., in the case of a single vacuum tube with the grid circuit in the jth mesh and the plate circuit in the kth mesh, as shown in Fig. 1. Fig. 2 shows the equivalent circuit of the tube; the input impedance Z_g of the tube is in the jth mesh and the plate impedance and a generator of value $-\mu e$ are in the kth mesh. e designates the grid voltage on the tube.

In this particular instance the jth mesh equation is

$$
Z_{j1}I_1 + Z_{j2}I_2 + \ldots + (Z_{jk}+0)I_k + \ldots + Z_{jn}I_n = 0 \quad (2A)
$$

where the kth mesh equation is

$$
Z_{k1}I_1 + Z_{k2}I_2 + \ldots + Z_{kn}I_n = -\mu Z_g I_j
$$

Thus, equivalently

$$
Z_{k1}I_1 + Z_{k2}I_2 + \ldots + (Z_{kj}+\mu Z_g)I_j + \ldots + Z_{kn}I_n = 0 \quad (2B)
$$

The unilateral property of the tube is evidenced by the fact that in equations 2(A) and (B)

$$
Z_{kj}' = (Z_{kj}+\mu Z_g) \text{ does not equal}
$$
$$
Z_{jk}' = (Z_{jk}+0) \quad (2C)
$$

Thus, the reciprocal property does not accrue to the primed impedances Z_{kj}' and Z_{jk}'.

If now in an arbitrarily specified system a similar transformation is effected for each of the unilateral elements in turn, the resulting set of equations will be of the form

$$
\begin{aligned}
Z_{11}I_1 + Z_{12}I_2 + \ldots + Z_{1n}I_n &= E_1 \\
Z_{21}I_1 + Z_{22}I_2 + \ldots + Z_{2n}I_n &= 0 \\
&\vdots \\
Z_{n1}I_1 + Z_{n2}I_2 + \ldots + Z_{nn}I_n &= 0
\end{aligned} \tag{3}
$$

where now the Z_{ij} in equation 3 may represent impedances such as those of Z_{kj}' and Z_{jk}' in equation 2, the primes being dropped for convenience. Inasmuch as $Z_{ij} \neq Z_{ji}$ in general, the determinant comprised of the coefficients of equation 3 is, in general, not symmetrical about the main diagonal.

By well-known algebraic theory, solution of equation 3 for the mesh current I_1 in response to the driving voltage E_1 is

$$
I_1 = \frac{\Delta_{11}}{\Delta} E_1 \tag{4}
$$

where Δ designates the determinant of the coefficients of equation 3 and Δ_{11} designates the first cofactor obtained by omitting the first row and column of Δ. Similarly

$$
I_2 = \frac{\Delta_{12}}{\Delta} E_1
$$

$$
I_3 = \frac{\Delta_{13}}{\Delta} E_1
$$

$$
\vdots \tag{5}
$$

$$
I_n = \frac{\Delta_{1n}}{\Delta} E_1
$$

RETURN-VOLTAGE DIFFERENCE AND RETURN RATIO

In the preceding subsection the impedances (or admittances when nodal equations are used) of an active element or passive unilateral coupling device were encompassed in the terms of the basic determinant Δ. A method of measuring feedback is now given. Consider the circuit shown in Fig. 3. This figure represents a single tube in a multiple-loop system. The block N represents the complete circuit except for the tube in question; P_1

Fig. 2 (below). Equivalent circuit for a triode

Fig. 3 (right). System when measuring voltage return difference
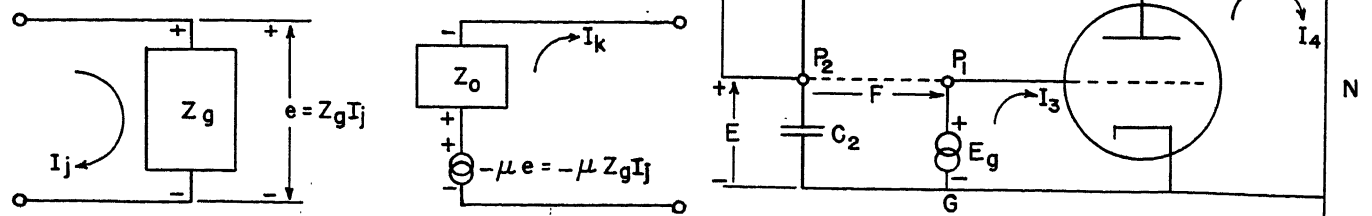
and P_2 are points that are connected together in normal operation. Consider the indicated generator of driving voltage $e_g(t) = \sin \omega t$; transformation yields $E_g(s) = \omega/(s^2+\omega^2)$. Note that C_1 and C_2, representing capacitances equal to the plate-to-grid and grid-to-cathode capacitance, are shown connected to P_2; hence the circuit structure is not disturbed when the connection between P_1 and P_2 is opened, as indicated by the dotted line.

The return-voltage difference F is defined as the phasor representation of the steady-state alternating-voltage difference existing between P_1 and P_2 when the sinusoidal voltage characterized by $e_g = \sin \omega t$ is applied between P_1 and G. Thus, $F(j\omega) = [1 - E(j\omega)]$, per Fig. 3. The return ratio T is defined as the return-voltage difference minus 1; thus, $T = F - 1 =$ return ratio; hence $T(j\omega) = F(j\omega) - 1$.

Expression of F and T in terms of complex circuit impedances and a method of calculation of T can now be obtained. Let the grid terminal and plate terminal of the tube under consideration be respectively in, typically, meshes 3 and 4; and take, as occasion demands, the operational transconductance or mutual admittance of the tube as $W(s)$. W is thus a constituent of Z_{43} or Y_{43} in the general system of mesh or nodal equations respectively.

A current i_4 in the direction of the arrow at the point 4 will result in a voltage rise

$$E(s) = \frac{\Delta_{43}(s)}{\Delta^0(s)} I_4(s) \tag{6}$$

at point P_2, where Δ^0 represents the determinantal value with the tube dead $(W=0)$. The operational voltage $E_g(s) = \omega/(s^2+\omega^2)$ applied between 0 and P_1 will produce a current $I_4 = -W\omega/(s^2+\omega^2)$ in the direction of the arrow. The return-voltage difference $[E_g(j\omega) - E(j\omega)]$ is calculable as the phasor value of the steady-state component stemming from the transform of the voltage difference $[E_g(s) - E(s)]$ between P_1 and P_2. Thus

$$F = \left\{ L^{-1}\left[\frac{\omega}{s^2+\omega^2} \times \left(1 + W\frac{\Delta_{43}}{\Delta^0} \right) \right] \right\}_{ss \text{ (in phasor form)}} \tag{7A}$$

$$T = F - 1 \tag{7B}$$

Performing the inversion, selecting the steady-state sinusoidal terms, and expressing these as phasors yields

$$F = 1 + W(j\omega)\frac{\Delta_{43}(j\omega)}{\Delta^0(j\omega)} \tag{7C}$$

$$T = W\frac{\Delta_{43}(j\omega)}{\Delta^0(j\omega)} \tag{7D}$$

where it is noted that the determinants are now functions of $j\omega$.

It follows from the properties of determinants that if $\Delta^0(j\omega)$ represents the value of the determinant Δ when $W=0$ and $W(j\omega)$ enters into the determinant only in $Z_{43}(j\omega)$, then $\Delta^0(j\omega) + W\Delta_{43}(j\omega)$ represents the value of the determinant $\Delta(j\omega)$ when W has its normal value. Hence

$$F = \frac{\Delta(j\omega)}{\Delta^0(j\omega)} \tag{8}$$

The question of the stability of the general multiple-loop structure can now be treated.

DETERMINATION OF STABILITY

According to equation 5 and the theory developed in Section 1 and subsequent sections, the response $i_y(t)$ of the general structure is stable if, when $e_1(t) = \delta(t)$ and thus $E_1(s) = 1$, the roots of the determinantal equation $\Delta(s) = 0$ have negative real parts. Limited stability occurs if some roots have negative real parts and the remaining roots are zero or imaginary, and the zero root is single and the imaginary roots are either simple, or, if they are not simple, an appropriate number of minors have this same zero. Detailed consideration of limited stability is treated in a subsequent section; now only determination of the zeros of $\Delta(s)$ other than on the imaginary axis need be considered.

If the number of zeros of $\Delta(s)$ is obtained through a plot of the Nyquist diagram of the return difference given in equation 8 for values of $s = j\omega$ as ω runs from $\omega = -\infty$ to $\omega = +\infty$ and s then runs along the right-hand infinite semicircle from $s = j\infty$ to $-j\infty$, the information obtained may be ambiguous; for according to Cauchy's principle of argument the algebraically counted number of times the plot in the $F(s)$ plane encircles the origin in a counterclockwise direction gives the difference between the number of zeros of Δ^0 and of Δ in the right half-plane of s. But this information is not particularly useful, inasmuch as Δ^0 may have zeros in the right half-plane. The single-loop case gives no such difficulty, because Δ^0 then represents the circuit determinant of a passive system (the one and only active element is dead) whence $\Delta^0 = 1$. The Δ^0 of a multiple-loop system may, however, have zeros in the right half-plane because of the presence of the active elements, except the last one, being in their normal conditions.

In such case the procedure to be used is to assume all the active elements initially dead and plot the return differences

as the active elements are restored one at a time to their normal condition. For it follows from equation 8 that the Nyquist diagram when all the tubes up to and including the jth tube are operative will encircle the origin $(P_j - Z_j)$ times in a counterclockwise direction, where P_j and Z_j respectively represent the number of poles and zeros of $F_j = \Delta_j/\Delta_j^0$ in the right half-plane. The total number of encirclements is thus $(P_1 - Z_1) + (P_2 - Z_2) + \ldots + (P_n - Z_n)$. The denominator of the jth return difference is the numerator of the $(j-1)$th return difference. Therefore $P_j = Z_{j-1}$. Thus the total number of encirclements equals $(P_1 - Z_n)$. Now $P_1 = 0$, because when the system comprises only bilateral passive elements, all the active elements being inoperative, Δ_1^0, the resulting system determinant, has only zeros with negative real parts. Hence the total number of encirclements is Z_n. Thus, if the total number of algebraically counted counterclockwise encirclements of the series of Nyquist diagrams obtained by beginning with all the active elements dead and energizing them one by one is equal to zero, the final system obtained is stable; if not, the system is unstable.

It should be noted that the order in which the elements are re-energized does not affect the end result; however, it may well be that one order of restoration yields simpler intermediate computation than another. If so, and if the simplest order of restoration is discernible, it is obviously advantageous to so proceed.

GENERALIZATION TO ENCOMPASS LIMITEDLY STABLE NETWORKS

The method just described for ascertaining stability is based on the assumption that none of the determinants Δ_i or Δ_i^0 occurring in the return difference F_i lie on the imaginary axis of a plane. This restriction will now be removed, and a theory formulated enabling the calculation of limitedly stable systems.

The occurrence on this axis of s of a zero of a determinant causes, first, a zero of the return difference having this determinant in the numerator, and second, a pole in the succeeding return difference, where this determinant occurs in the denominator. The plots of these return differences then go respectively through zero and to infinity. At these points and in so far as the one determinant is concerned, abrupt changes in the arguments of the corresponding return differences occur. The values of the changes are given by the use of a generalization of Nyquist's method effected by application of the generalized principle of argument stated in Appendix I. Thus a sim-
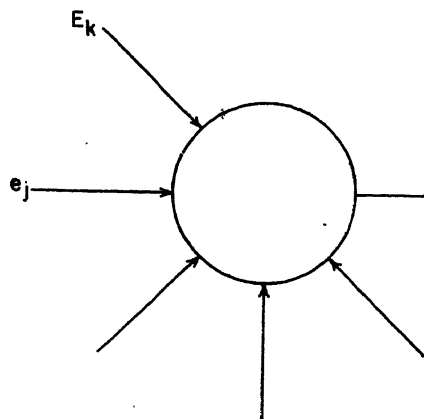
Fig. 4 (left). Schematic of a unilateral comparator

Fig. 5 (right). Schematic of an active element
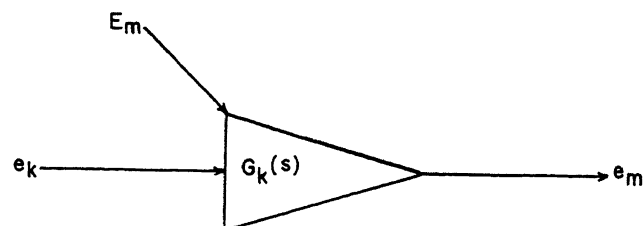
ple zero results in a change of π, minus or plus according as the zero of the determinant is a zero or a pole of the return difference; a double zero results in a change of 2π; a triple zero in a change of 3π, etc.

Two major cases must be treated: one of the intermediate determinants Δ_i ($i \neq n$) may have zeros on the imaginary axis of s; or the characteristic determinant Δ representing the complete structure may have such zeros.

Case Where Imaginary Zeros Occur in an Intermediate Determinant

The condition where an intermediate determinant has zeros on the imaginary axis of s need be treated only to the end that the number of encirclements of the corresponding return difference be known without ambiguity. An intermediate determinant occurs once as a denominator and once as a numerator in the total sequence of return differences. Accordingly the discontinuity of argument is of opposite sign in the two associated plots of the F. Thus, on summing the total changes in argument of all the plots, the net change of argument due to these particular zeros will be zero. In such case the net encirclements for all the plots will be as though these particular zeros had not occurred.

Case Where Imaginary Zeros Occur in the Last Determinant

Next the behavior of the structure is considered if active elements fail so that the intermediate determinant which has zeros on the imaginary axis of s is the characteristic determinant of the system. In such case it follows from Section 1 and subsequent sections that in order to insure limited stability the zeros occurring on the imaginary axis must be simple or else an appropriate number of the minors must vanish for these same zeros. Analysis of this case evidently parallels that for the case when the determinant representing the complete structure has

zeros on the imaginary axis. It therefore suffices to treat this final case.

Assume then that the Nyquist plot of the last return difference F goes through $F(s) = 0$. Since the plot is for values of s on the imaginary axis of the s-plane, this indicates that $F(s)$ has zeros thereon. This function F is the quotient of two determinants whose expansions are polynomials in s. The numerator Δ is the characteristic polynomial representing the system. If $F(s)$ vanishes, so also must $\Delta(s)$; hence it has zeros on the imaginary axis of s. However, this fact necessitates no essential change in the procedure outlined. The only difference will be that in ascertaining $(P_n - Z_n)$ from the plot of F_n, the calculation must be according to the generalized Nyquist method given in the second part of the Appendix rather than the usual simpler Nyquist criterion of the first part of the Appendix.

The analysis for zeros of one or more of the Δ_j occurring on the imaginary axis is now completed, which, in turn, marks the end of a total analysis enabling complete ascertainment of the stability of a general linear lumped-parameter multiple-loop network encompassing both active and passive elements.

6. General Stability Analysis of a Linear Lumped-Parameter Multiple-Loop Servomechanism

The complete stability analysis of linear lumped-parameter multiple-loop active networks detailed in Section 5, applicable on either the mesh or nodal approach, is based on the use of operational impedances or admittances. Use of the latter is suited to feedback amplifier analysis because the equations are most easily expressed in terms of impedances or admittances. Currently, however, servomechanism theory is largely based on the use of transfer functions. Accordingly, it is desirable to develop a general theory akin to that of Section 5, but couched in terms of transfer functions rather than impedances or admittances.

Equations Characterizing Comparison Device and Active and Passive Elements

A comparison element (which may be a pair of synchros, a differential trans-

former, an adding amplifier, or any such device) of a servo system is shown schematically in Fig. 4; this element is commonly termed a unilateral comparator. Working now in terms of the Laplace transforms of the system quantities, let the signals e_j and e_k be assumed as output and input signals of the active elements; the input E_k of Fig. 4 represent some externally introduced signal. The action of the comparator is such that the quantity e_k is equal to e_k', the signal that would exist if no input were introduced, plus the input E_k; i.e., $e_k = e_k' + E_k$. The unidentified arrows leading into the lower part of the comparator represent inputs due to signals in other parts of the structure, which come through passive networks to this comparator, in accordance with a convention that if inputs come from an active element they will be designated by subscripts, as are e_j and e_k.

The equation describing the function of the comparator can be written as

$$e_1 P_{k1} + e_2 P_{k2} + \ldots + e_j P_{kj} + e_k P_{kk} + \ldots + e_n P_{kn} = E_k P_{kk} \quad (9)$$

P_{kk} will in general be equal to 1. This equation could be solved for e_k, which might nominally be taken as the output of the comparator. This done, $(-e_1 P_{k1}/P_{kk})$ can be interpreted as the component of e_k due to e_1, $(-e_2 P_{k2}/P_{kk})$ as the component due to e_2, etc. Correspondingly, $(-P_{k1}/P_{kk})$ can be interpreted as the transfer function from e_1 to e_k, etc.

An active element, e.g., an amplifier (electronic, hydraulic, pneumatic, etc.) is shown schematically in Fig. 5. Here E_m represents the externally introduced signal. The output signal e_m is given by

$$e_m = G_k(s) e_k + E_m = e_m' + E_m \quad (10)$$

A passive element will have associated with it an equation similar to equation 10, except that in general it will contain no externally introduced signal.

General System Equations

If for a quiescent system an equation such as 9 is used for each comparator, and equations similar to equation 10 for each active element and each passive element, and all the e_j are eliminated except those which are input or output signals of the active elements, a set of equations is obtained, typified by

$$e_1P_{11}+e_2P_{22}+\ldots+e_nP_{1n}=E_1P_{11}$$
$$e_1P_{21}+e_2P_{22}+\ldots+e_nP_{2n}=E_2P_{22}$$
$$\vdots$$
$$e_1P_{n1}+e_2P_{n2}+\ldots+e_nP_{nn}=E_nP_{nn}$$

(11)

In equation 11 all symbols designate the Laplace transforms of the corresponding system quantities, hence are functions of s.

RETURN-SIGNAL DIFFERENCES IN TERMS OF TRANSFER FUNCTIONS

The procedure for ascertaining stability given in Section 5 requires plotting the return-voltage differences. Essentially, this approach enables the testing of stability because the sequence of return-voltage differences terminates in a final return-voltage difference whose numerator is the characteristic determinant of the system. A similar procedure is used here.

Consider an active element, say the second. It is desired to find the return-signal difference. The system of equation 11 defines the system when the points P_1 and P_2 in Fig. 6 are connected together and the test generator shown by $e_g(s)$ is removed. For this condition, find the voltage e_2 due to the singly acting external voltage E_3; all the other E_j are assumed to be zero.

By the way in which the equations are formed, $G_2(s)$ appears only once in the system of equation 11; namely in

$$-e_2G_2(s)+e_3=E_3$$

(12)

Under these conditions equation 11 becomes

$$e_1P_{11}+e_2P_{12}+e_3P_{13}+\ldots+e_nP_{1n}=0$$
$$e_1P_{21}+e_2P_{22}+e_3P_{23}+\ldots+e_nP_{2n}=0$$
$$0-e_2G_2(s)+e_3+\ldots+0=E_3$$
$$\vdots$$
$$e_1P_{n1}+e_2P_{n2}+e_3P_{n3}+\ldots+e_nP_{nn}=0$$

(13)

Solution of this set for e_2 yields

$$e_2=E_3\frac{\Delta_{32}}{\Delta}$$

(14)

When the system is as in Fig. 6, the active element is inoperative in the system except that its output $G_2(s)e_g$ acts to provide an external signal source equivalent to the E_3 utilized in the foregoing. With $G_2(s)=0$ and a signal E_3 applied, equation 15 is obtained from equation 14

$$e_2=E_3\frac{\Delta_{32}}{\Delta^0}$$

(15)

where Δ^0 indicates the value of Δ with $G_2(s)=0$, and Δ_{32} remains unchanged inasmuch as $G_2(s)$ does not occur therein (see the remark preceding equation 12). If $G_2(s)e_g$ is treated as the signal source equivalent to E_3, then

$$e_2=e_gG_2(s)\frac{\Delta_{32}}{\Delta^0}$$

(16)

The return-signal difference is given by e_g-e_2, whence by equation 16

$$F=e_g-\left[e_gG_2(s)\frac{\Delta_{32}}{\Delta^0}\right]$$
$$=e_g\left[1-G_2(s)\frac{\Delta_{32}}{\Delta^0}\right]$$

(17)

$$=e_g\left[\frac{\Delta^0-G_2(s)\Delta_{32}}{\Delta^0}\right]$$

But Δ^0 is evaluated with $G_2(s)=0$, and $G_2(s)$ occurs only as $-P_{32}$. As indicated in equation 13 the third row of Δ contains only two elements, P_{32} and 1. Hence Δ can be written as the sum of the two determinants

$$\Delta^0-G_2(s)\Delta_{32}=\Delta$$

(18)

Substituting yields

$$F(s)=\frac{\Delta}{\Delta^0}e_g(s)$$

(19)

Taking $e_g(s)=1$, which corresponds to a unit impulse test signal $e_g(t)=\delta(t)$

$$F(s)=\frac{\Delta(s)}{\Delta^0(s)}$$

(20)

If, alternately, $e_g(t)=\sin\omega t$, then $e_g(s)=\omega/(s^2+\omega^2)$ and $F(s)=[\omega/(s^2+\omega^2)][\Delta(s)/\Delta^0(s)]$. Inversion of $F(s)$, and selection of the sinusoidal steady-state component of the resulting time function and expression in phasor form as in the second part of Section 5, yields

$$F(j\omega)=\frac{\Delta(j\omega)}{\Delta^0(j\omega)}$$

(21)

Obviously, this complex number form of the return-signal difference enables lab-

oratory determination of the return-signal differences. These $F(j\omega)$ are precisely those yielded by equation 20 as s runs over the imaginary axis.

As equation 21 is of the same functional form as equation 8, it follows that determination of the stability of the system characterized by the set of equations 11 can be carried out in precisely the same manner as outlined in the third part of Section 5. In the course of this analysis it is necessary to calculate the signal-return differences $F_j=\Delta_j/\Delta_j{}^0=\Delta_j/\Delta_{j-1}$ for $j=1,\ldots,m$ (where m designates the number of active elements). Here F_j designates the return difference obtained with $e_g(t)=\sin\omega t$ applied at the input of $G_j(s)$, with the active elements $G_{j+1},\ldots,$ G_m inoperative: i.e., $G_{j+1}(s)=\ldots=G_m(s)=0$. Accordingly, the set of signal-return differences is needed

$$F_j(j\omega)=\frac{\Delta_j(j\omega)}{\Delta_{j-1}(j\omega)}$$

(21A)

where $\Delta_j(j\omega)=[\Delta(s)$ evaluated for $G_{j+1}(s)=\ldots=G_m(s)=0]_{s=j\omega}$. With this knowledge it is possible to ascertain stability through the use of Nyquist's stability method, which is illustrated now for a specific system.

ILLUSTRATIVE EXAMPLE

This example is sufficiently broad in scope to illustrate use of the whole of the theory just developed. Fig. 7 shows a multiple-loop system. The system of equations paralleling equation 11 can be written by inspection. Thus
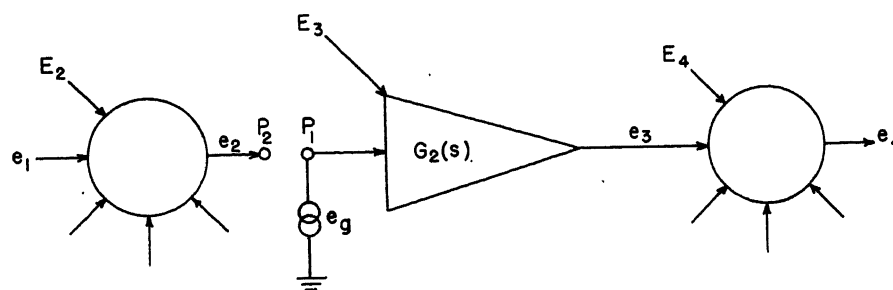
Fig. 6. System when calculating return difference
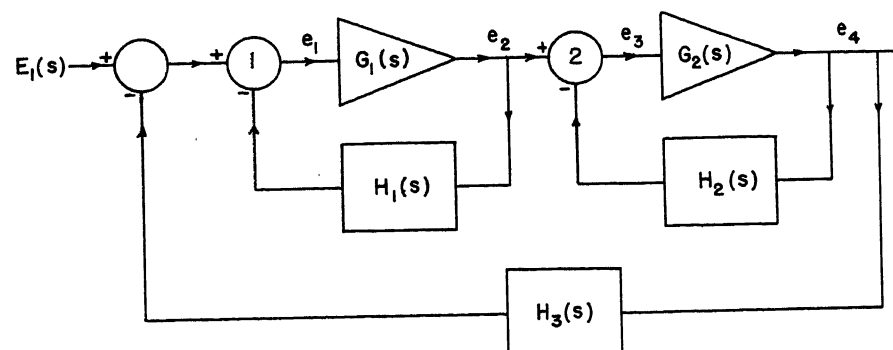
Fig. 7. Illustrative multiple-loop system
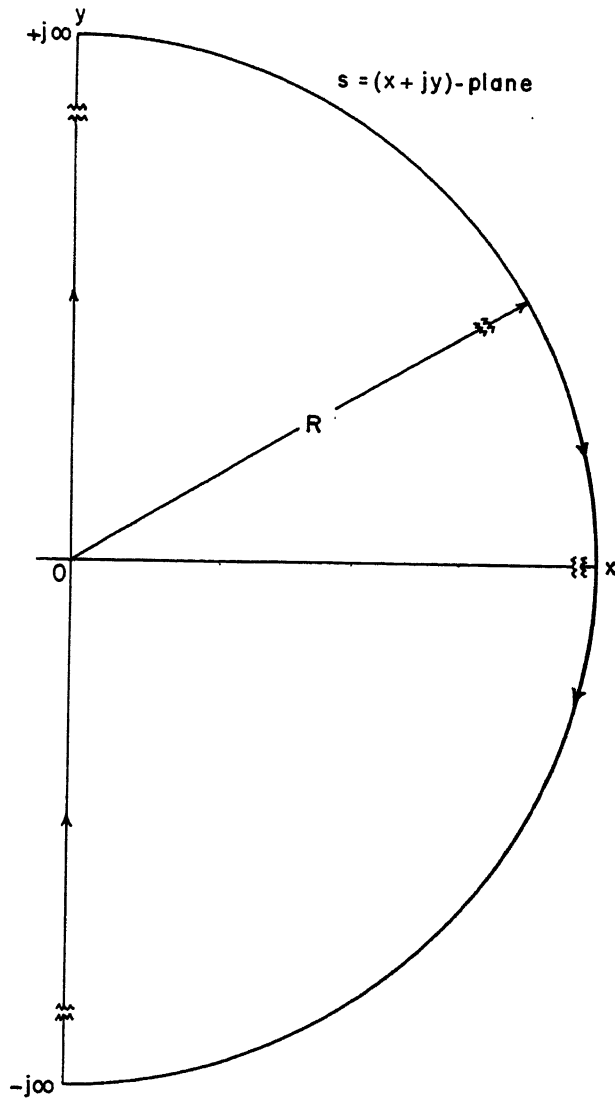
Fig. 8. Contour bounding the right half-plane of s; no poles on the imaginary axis

Fig. 10. Solid line is plot of F_2 for the first case; dashed line is a detailed plot in vicinity of the origin to an enlarged scale

First comparator:

$$e_1 + H_1(s)e_2 + 0 + H_3(s)e_4 = E_1$$

First amplifier:

$$-G_1(s)e_1 + e_2 + 0 + 0 = -0 \quad (22)$$

Second comparator:

$$0 - e_2 + e_3 + H_2(s)e_4 = 0$$

Second amplifier:

$$0 + 0 - G_2(s)e_3 + e_4 = 0$$

Equations 22 can be reduced by inspection

Fig. 9. Plot of F_1 for the first case

to

$$[1 + G_1(s)H_1(s)]e_1 + [G_2(s)H_3(s)]e_3 = E_1$$
$$[-G_1(s)]e_1 + [1 + G_2(s)H_2(s)]e_3 = 0 \quad (23)$$

Solution of equation 23 for e_3 yields

$$e_3 = \frac{G_1(s)E_1}{1 + H_1(s)G_1(s) + H_2(s)G_2(s) + H_1(s)H_2(s)}$$
$$G_1(s)G_2(s) + G_1(s)G_2(s)H_3(s) \quad (24)$$

From

$$e_4 = G_2(s)e_3 \quad (25)$$

equation 26 is obtained

$$e_4 = \frac{G_1(s)G_2(s)E_1}{1 + H_1(s)G_1(s) + H_2(s)G_2(s) +}$$
$$H_1(s)H_2(s)G_1(s)G_2(s) + G_1(s)G_2(s)H_3(s) \quad (26)$$

Imposing the condition $G_2 = 0$ yields $e_4 = 0$; hence $e_2 = e_3$ for this case. The over-all transfer function of the first loop with the second amplifier inoperative is obviously

$$\frac{e_2}{E_1} = \frac{G_1(s)}{1 + H_1(s)G_1(s)} \quad (27)$$

The determinant of the coefficients of the set of equations 23 is

$$\Delta = \begin{vmatrix} [1 + G_1(s)H_1(s)] & [G_2(s)H_3(s)] \\ [-G_1(s)] & [1 + G_2(s)H_2(s)] \end{vmatrix} \quad (28)$$

Calculating Δ_0, Δ_1, and Δ_2 [the values of Δ found by taking, in turn, $G_1(s)$ and $G_2(s)$ equal zero, hence both amplifiers in Fig.
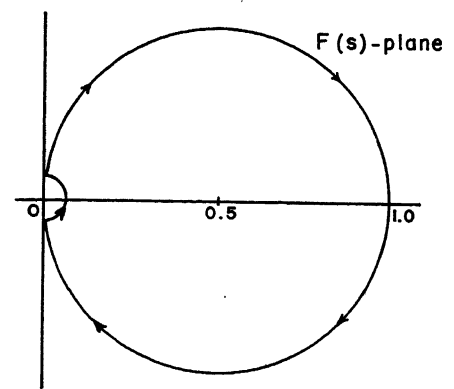
Fig. 11. Plot of F_1 for the second case

F (s) - plane

Fig. 12. The plot goes to infinity twice, indicating a pair of imaginary poles. By the generalized argument principle, the net change in the phase of F_2 as y runs over the successive segments of the imaginary axis is 2π, equivalent to one counterclockwise encirclement.

Accordingly, the net sum of the encirclements of F_1 and F_2 is zero encirclements, whence from $0=P_1-Z_2=0-Z_2$, we have $Z_2=0$ and the system is stable. In this instance, if the amplifier of transfer function G_2 becomes inoperative the net encirclements are given by F_1 alone. Hence $-1=P_1=Z_1=0-Z_1$, or $Z_1=1$. This corresponds to either one zero of F_1 in the right half-plane or two on the imaginary axis. Inspection of Fig. 11 shows that the latter is the case. Thus the partially functioning system is limitedly stable.

Inner Loop Unstable, Outer Loop Limitedly Stable: K=3.309

From equation 30

$$F_1=\frac{s^2-1.309s+1}{s^2+2s+1}$$

Fig. 13 shows the Nyquist plot of F_1. As before, the curve representing $+y$ and that representing $-y$ coincide. The total change of argument for F_1 is -4π, corresponding to two clockwise encirclements. Thus $P_1-Z_1=-2$, but since $P_1=0$ we have $Z_1=2$. From equation 31

$$F_2=\frac{s^4+5.691s^3+9.382s^2+38.8s+17.55}{(s^2+2s+1)(s^2-1.309s+1)}$$

Fig. 14 displays the Nyquist plot of F_2. Application of the argument principle indicates a change of phase of 2π, corresponding to one counterclockwise encirclement. Thus $P_1-Z_2=0-Z_2=1$, indicating either a zero in the right half-plane or two conjugate imaginary zeros. Inspection of Fig. 14 shows the latter to be the case. Accordingly, the system is limitedly stable.

AN ALTERNATIVE APPROACH

Rather than ascertain the difference between poles and zeros of the F_i by use of the generalized principle of argument, a conformal mapping procedure could be used. For details and illustration

---

7 inoperative; $G_2(s)=0$, hence, the second amplifier in Fig. 7 inoperative; and both amplifiers operative] from equation 28 yields

$$\Delta_0=1$$
$$\Delta_1=1+G_1(s)H_2(s) \qquad (29)$$
$$\Delta_2=1+G_1(s)H_1(s)+G_2(s)H_2(s)+$$
$$G_1(s)G_2(s)H_1(s)H_2(s)+G_1(s)G_2(s)H_3(s)$$

where, from use of equation 21(A)

$$F_1=\frac{1+G_1(s)H_1(s)}{1} \qquad (30)$$

$$F_2=\frac{1+G_1(s)H_1(s)+G_2(s)H_2(s)+G_1(s)\times}{1+G_1(s)H_1(s)} \frac{G_2(s)H_1(s)H_2(s)+G_1(s)G_2(s)H_3(s)}{}$$
$$\qquad (31)$$

These functions, when plotted as s runs over the contour of Fig. 8, enable the determination of stability as detailed previously. As a particular case, consider a system with transfer functions

$$G_1(s)=\frac{K}{s+1} \qquad H_1(s)=\frac{-s}{s+1} \qquad H_3(s)=1$$

$$G_2(s)=\frac{5}{s+1} \qquad H_2(s)=\frac{s}{s+1} \qquad (32)$$

While the values of the gain constants and time constants thus chosen are not particularly representative of those occurring in practice, a choice of integral values avoids obscuring the run of the development by numerical data not easily handled.

Stability is now investigated for three different values of K ($K=3.0$, 2.0, and 3.309), so chosen as to illustrate all points involved in stability calculation by the theory derived in the foregoing.

Inner Loop Unstable, Complete Structure Stable: K = 3.0

From equation 30

$$F_1=\frac{1+G_1(s)H_1(s)}{1}=\frac{s^2-s+1}{s^2+2s+1}=\frac{\Delta_1}{\Delta_0} \qquad (33)$$

The denominator of F_1 has no zeros in the right half-plane, hence $P_1=0$. Fig. 9, the Nyquist plot of F_1, where the portions of the plot for positive and negative values of y are coincident, indicates two clockwise encirclements. From equation 31

$$F_2=\frac{s^4+6s^3+10s^2+36s+16}{(s^2+2s+1)(s^2-s+1)} \qquad (34)$$

Fig. 10, the Nyquist plot of this function, indicates two counterclockwise encirclements.

Accordingly, the net sum of the encirclements of F_1 and F_2 is zero; whence from $0=P_1-Z_2=0-Z_2$, we have $Z_2=0$ and thus the system of equation 26 is stable. Suppose, however, the amplifier having the transfer function $G_2(s)$ became inoperative in practice. Then the net encirclements are given by the plot of F_1 alone. As then, from $-2=0-Z_1$, we have $Z_1=2$, the partially functioning system of equation 27 would be unstable.

Inner Loop Limitedly Stable, Outer Loop Stable: K=2

From equation 30

$$F_1=\frac{s^2+1}{s^2+2s+1}$$

The denominator of F_1 has no zeros in the right half-plane, hence $P_1=0$. Fig. 11 shows the Nyquist plot of F_1. Here, as in Fig. 10, the portions for y and $-y$ coincide. The plot goes through zero twice. This indicates a pair of conjugate imaginary zeros. By the generalized principle of argument, the net change in phase of $F_1(jy)$ as y runs over the successive segments of the imaginary axis is -2π, equivalent to one clockwise encirclement. From equation 31

$$F_2=\frac{s^4+7s^3+12s^2+27s+11}{(s+1)^2(s^2+1)}$$

The Nyquist diagram of F_2 is plotted in

Fig. 13 (left). Plot of F₁ for the third case

Fig. 15 (right) (A). The function $f(s)$ is analytic on C and is assumed to possess no zeros on C. (B). The contour C is indented to the inside of the enclosed area to by-pass a zero of $f(s)$ at the origin and poles of $f(s)$ at $s = \pm ja$

A   B

through solution of the illustrative problem of Fig. 7, see reference 3.

Appendix I. Methods of Nyquist and Cauchy

Nyquist's Method and Cauchy's Principle of Argument

Nyquist's[5] original derivation was limited to considering the characteristic equation of a system whose output for sinusoidal input vanished as the frequency approached infinity; he obtained his final expression by a rather intricate analysis. However, both the restriction and the intricacy of analysis can be obviated by a derivation based on Cauchy's principle of argument, easily established by simple complex variable theory.

Theorem I: Let $f(s)$ be single-valued and analytic both inside and on the simple closed contour C bounding the domain D, except for a possible number p of discrete poles of

multiplicity m_i at s_i. Let $f(s) \neq 0$ anywhere on C. Then

$$\frac{1}{2\pi i}\oint_C \frac{f'(s)}{f(s)}\,ds = P - Z = \frac{1}{2\pi}\Delta[\text{phase } f(s)]$$

where P designates the number of poles $P = \sum_{i=1}^{p} m_i$; Z designates the number of zeros $Z = \sum_{j=1}^{z} n_i$ of $f(s)$ inside C; and $\Delta[\text{phase } f(s)]$ indicates the algebraic increase of phase (argument) of $f(s)$ as the contour C is traversed once in the evidenced clockwise* direction, phase being valued positive or negative in the usual trigonometric sense. Proofs of this theorem are advanced in numerous textbooks on complex variable theory; also, specifically, in references 3 and 6. In this light Nyquist's criterion as utilized in servomechanism theory is merely an adoption of Cauchy's principle of argument to the contour of Fig. 8, where $f(s)$ is

---

*According to established convention in servomechanism analysis, integration along the contour of Fig. 8 is in the opposite direction to the conventional mathematically positive sense used in the theorem; thus the direction of integration has been chosen in accordance with this servomechanism convention.

to be interpreted as the appropriate characteristic polynomial or transfer function of a servo system.

Generalization of Nyquist's Method and Cauchy's Principle of Argument

In multiple-loop servomechanisms it occasionally happens that zeros of the various characteristic functions occur in general on the imaginary axis; and in both multiple-loop and single-loop analysis, a zero often occurs at the origin. Obviously, therefore, it is desirable to so generalize Nyquist's criterion that calculation can be effected for such cases. This generalization is obtained through use of a generalization of Cauchy's principle of argument as follows:

Theorem II: Let the conditions on $f(s)$ be as stated in theorem I, except that $f(s)$ may have possible discrete poles and zeros on C. Let C_d be the contour obtained by indenting C with semicircles of indefinitely small radius as in Fig. 15, so as to by-pass the poles and zeros of $f(s)$ on C. Then

$$\frac{1}{2\pi j}\oint_C \frac{f'(s)}{f(s)}\,ds = P - Z = \frac{1}{2\pi}\times$$

$$\Delta_s[\text{phase of } f(s)] - \frac{P'}{2} + \frac{Z'}{2}$$

where P and Z are as in theorem I; P' designates the number of "indented" poles $P' = \sum_{i=1}^{p'} m_i'$; Z' designates the number of "indented" zeros $Z' = \sum_{j=1}^{z'} n'_j$; and Δ_s [phase of $f(s)$] indicates the algebraic increase of phase (argument) of $f(s)$ on the separate segments of C_d which link the indenting semicircles.

Proofs of theorem II are advanced in references 3, 6, and 7. In this paper the generalization of Nyquist's criterion comprises adoption of theorem II in the same fashion as outlined for theorem I. The plot of the Nyquist diagram yields evaluation of Δ_s [phase of $f(s)$], and P' and Z' are shown *per se;* thus knowledge of $(P - Z)$ is obtained as desired.

Fig. 14. Solid line is plot of F₂ for the third case; dashed line is a detailed plot in vicinity of the origin to an enlarged scale

References

1. The Current Status of Dynamic Stability Theory, Frank E. Bothwell. *AIEE Transactions,* vol. 71, pt. I, July 1952, pp. 223–28.

2. Network Analysis and Feedback Amplifier

DESIGN (book), H. W. Bode. D. Van Nostrand Company, Inc., New York, N. Y., 1948.

3. GENERALIZED STABILITY CRITERIA FOR LINEAR MULTIPLE-LOOP SERVOMECHANISMS, Thomas S. Amlie. *Ph.D. Thesis*, University of Wisconsin, Madison, Wis., 1952. See Appendixes A and B.

4. ADVANCED PART OF A TREATISE ON ADVANCED RIGID DYNAMICS (book), E. J. Routh. Cambridge University Press, Cambridge, England, 6th ed., 1930, pp. 210–31.

5. REGENERATION THEORY, H. Nyquist. *Bell System Technical Journal*, New York, N. Y., vol. 11, 1932, pp. 126–47.

6. CONFORMAL MAPPING (book), Z. Nehari. McGraw-Hill Book Company, Inc., New York, N. Y., 1952.

7. A NEW SIMPLE GENERAL CRITERION FOR DETERMINING THE STABILITY OF SERVOMECHANISMS, T. J., E. M. Higgins, Michalak. *Proceedings*, Tenth National Electronics Conference, Chicago, Ill., Oct. 1954.

———————◆———————

Discussion

L. F. Kazda (University of Michigan, Ann Arbor, Mich.): I would like to compliment Dr. Amlie and Dr. Higgins on their clear and concise delineation of Bode's discussion of return ratio and return-voltage difference. This section of Bode's book has been a topic for discussion among those interested in closed loops for some time. The authors have made a useful contribution in applying these feedback amplifier concepts to multiple-loop servomechanisms.

Herbert K. Weiss (Northrop Aircraft, Inc., Hawthorne, Calif.): The authors are to be commended for their comprehensive delineation of procedures for ascertaining the stability of linear lumped-parameter multiple-loop feedback systems. Of especial interest is their careful establishment of the definitions of stability, and the interpretation of these definitions in terms of the dynamic performance of the system. However, it is to be regretted that the authors chose to omit from the paper those system types which are less commonly treated in the literature yet whose stability can be examined by extensions of the methods presented. These include systems with distributed parameters and systems which can be described by linear differential equations with time varying coefficients.

Designers of ultrahigh performance control systems such as those employed in missiles have pointed out to me that stability determination of a system often occurs at a late stage in system synthesis, and that the major initial effort is concerned with exploration of the bounds beyond which system elements become nonlinear. Lifting surfaces stall, moveable elements encounter stops, amplifiers saturate, and power supplies are limited. Within these bounds the system must perform stably, as one element after another exceeds its bound of linearity and returns. In addition the transfer functions of the constituent elements change in time intervals which are sometimes comparable to the response times of the elements and the complete system.

This is not a criticism of the content of the paper under discussion, which is confined to the domain defined by its title. It is hoped, however, that future papers will extend the examination to less familiar varieties of linear controls and ultimately to systems of bounded linearity.

George J. Thaler (United States Naval Postgraduate School, Monterey, Calif.): Mr. Amlie and Professor Higgins are to be congratulated on a thorough and interesting analysis of a difficult and important subject. Their concise and complete general definition of stability is worth noting; the word "stability" has been used much too loosely in feedback control literature. The authors' application of the method of return-voltage differences is the first such application in the servo literature (as far as I know), and it is quite obvious that they have developed a convenient and satisfactory means for analyzing stability in multiple-loop systems. One is inclined to suspect that the method of return-voltage differences may have other applications in the field of servomechanisms and feedback control systems.

For those of us who prefer to derive our equations from the specific circuit diagrams, the return-voltage differences equations may be obtained readily without considering the general matrix equation. In Fig. 7 two return-voltage differences are required, F_1 and F_2, as specified by equations 30 and 31. The equation for F_1 may be derived as follows:

F_1 obtains when amplifier 2 is inoperative. The return voltage desired is that voltage V_1 appearing between comparator 1 and the input to amplifier 1 when the connection between these components is broken and a voltage $e_g(s) = 1$ is applied at the input of amplifier 1 when all other voltages zero. Then the return voltage difference is

$$F_1 = e_g(s) - V_1(s) = \frac{e_g(s) - V_1(s)}{e_g(s)}$$
$$\text{(for } e_g(s) = 1) \quad (35)$$

$$V_1(s) = -e_g(s) G_1(s) H_1(s) \quad (36)$$

and thus

$$F_1 = 1 + G_1(s) H_1(s) \quad (37)$$

In like manner F_2 is obtained at the input of amplifier 2

$$F_2 = \frac{e_g(s) - V_2(s)}{e_g(s)} \quad (38)$$

where $V_2(s)$ is the voltage appearing across a broken connection between comparator 2 and amplifier 2.

$$V_2(s) = -e_g(s) G_2(s) H_2(s) +$$
$$e_g(s) G_2(s) H_2(s) [-1] \frac{G_1(s)}{1 + G_1(s) H_1(s)}$$
$$= -e_g \frac{G_2(s) H_2(s) + G_1(s) G_2(s) H_1(s) \times H_2(s) + G_1(s) G_2(s) H_3(s)}{1 + G_1(s) H_1(s)}$$
$$(39)$$

from which

$$F_2 = 1 + \frac{G_2(s) H_2(s) + G_1(s) G_2(s) H_1(s) H_2(s) + G_1(s) G_2(s) H_3(s)}{1 + G_1(s) H_1(s)}$$
$$(40)$$

$$= \frac{1 + G_1(s) H_1(s) + G_2(s) H_2(s) + G_1(s) G_2(s) H_1(s) H_2(s) + G_1(s) G_2(s) H_3(s)}{1 + G_1(s) H_1(s)}$$

These equations are identical with equations 30 and 31.

Richard W. Jones (Northwestern University, Evanston, Ill.): The very thorough discussion of stability, both from the standpoint of definition and evaluation, contained in this paper should be a valuable aid in the analysis of complex systems. In particular, the notion that one is principally concerned with the behavior of a system following a disturbance is well emphasized. The authors have wisely referred to the work of Bode and to the immense store of information contained within that volume that bears directly upon feedback control problems.

The analytical basis for stability as stated in this paper makes use of the transformed variable s. This variable is usually related to the Laplace transform method of solution, and although the formulation of stability criteria in these terms is admittedly correct, it does tend to obscure the fact that stability is entirely a property of the differential equation and not of any method of solution. Thus Laplace transforms can be left completely out of this discussion, and the entire stability problem stated as an algebraic one related to the roots (or coefficients) of a polynomial. This modification in viewpoint is one that I believe has significant educational advantages.

In proposing the use of Bode's method of return differences for the assessment of multiple-loop systems it should not be inferred that the more conventional Nyquist plots of certain open-loop functions are inapplicable. In applying Cauchy's principle of argument the results can hardly be said to be ambiguous but rather that this test is not sufficient and that more information is required. That is, the mapping test itself yields only the value of the difference between the number of zeros and the number of poles within the contour, and since it is usually the number of zeros that is sought, the number of poles must be ascertained by other means. In the single-loop case this is readily done because the poles of the test function (the closed-loop equation) are also the poles of the open-loop transfer function and thus normally obtainable by inspection. On the other hand, with multiple loops the location of the open-loop poles is not evident and must be found by additional calculation. This is carried out by applying the same mapping criteria to each minor loop. In the case presented in the paper this would involve maps of

$$G_1' = \frac{G_1}{1 + G_1 H_1} \qquad G_2' = \frac{G_2}{1 + G_2 H_2} \quad (41)$$

These are of course most easily made by mapping the open-loop functions $G_1 H_1$ and $G_2 H_2$ and testing with respect to -1. Plots of the inverse transfer function will frequently expedite this process. The complete system is represented by

$$\frac{e_4}{E_1} = \frac{G_1' G_2'}{1 + G_1' G_2' H_3} \quad (42)$$

so that it may be mapped by plotting the function $G_1' G_2' H_3$ and again referring this to -1.

It will be noted that in terms of the foregoing notation

$$\frac{e_4}{E_1} = \frac{G_1'G_2'}{(1+G_1H_1)\left(\dfrac{1+G_1'G_2'H_3}{1+G_1H_1}\right)} = \frac{G_1'G_2'}{F_1F_2}$$

$$(43)$$

so that F_1 and F_2 are simply factors of the closed-loop equation for the entire system. However, each of these two factors involves sums as well as products, and it is doubtful if any significant saving in computational time is achieved by using them. The situation, stated roughly, appears to be that of plotting fewer, but relatively more complex, closed-loop functions as described in the paper, versus that of plotting more open-loop functions each of which is somewhat more readily calculated. The formulation of the transfer function e_4/E_1 (equation 43) enables one to see that the sum of the quantities $P-Z$ for all return differences is exactly that for the complete differential equation.

A final comment regarding the direction in which the contours on the complex plane are traced is believed to be in order. It is unfortunate that so much of the control literature has developed using the clockwise or negative sense in tracing the contour around the right-hand half-plane. Texts in mathematics almost exclusively present Cauchy's relations with the path traced in the counterclockwise direction, and there is no difficulty involved in adhering to this convention in any discussion of control theory. I would make a modest plea for a reconsideration of this matter to avoid the confusion that follows when students attempt to correlate control and complex function theory.

D. E. Zilmer (U. S. Naval Ordnance Test Station, China Lake, Calif.): This paper presents a noteworthy method for the stability analysis of linear lumped-parameter multiple-loop servomechanisms and other feedback systems. The method, which is based on Bode's method of return differences (ref. 2 of the paper), introduces a set of return differences, associated with a sequence of loops of the servomechanism, which are analogues of the return differences used by Bode in feedback amplifier theory. The stability of the sequence of inner loops and of the entire system then can be examined in a systematic manner by applying any one of the pertinent standard techniques (e.g., Nyquist's) to the set of return differences.

The first part of the paper is an excellent analysis of the stability of linear systems and is a valuable contribution to the literature of the theory. Included are a classification of the behavior of the system from the standpoint of stability and the important points of stability criteria and analyses. The presentation assembles logically related matter usually scattered throughout the literature.

J. J. Skiles (University of Wisconsin, Madison, Wis.): Although the Routh and Hurwitz stability methods do yield information of the yes or no variety relating to the question of stability, the superiority of the Nyquist stability method is well established because the Nyquist diagram gives additional information that is useful

in system design. Most books on servomechanism theory devote considerable attention to the use of the information obtainable from the Nyquist diagram of a single-loop system in the design of amplitude- and phase-compensating networks to improve system stability and/or frequency response. With but few exceptions, notably the book by MacColl,[1] only superficial treatment is given in textbooks to the Nyquist approach to the problem of the stability and compensation of multiple-loop systems. Thus, while the analysis and design of multiple-loop systems is admittedly more difficult than the design of single-loop systems, the limited use of the Nyquist approach to multiple-loop system design no doubt stems, at least in part, from the cursory treatment afforded this topic in textbooks.

The precise definitions of stability given in this paper and the lucid exposition of the Nyquist method for multiple-loop systems phrased in terms of return differences, return ratios, and transfer functions are welcome supplements to the less detailed and more restricted accounts set forth by Bode and MacColl.

REFERENCE

1. FUNDAMENTAL THEORY OF SERVOMECHANISMS (book), L. A. MacColl. D. Van Nostrand Company, Inc., New York, N. Y., 1945.

Thomas S. Amlie and Thomas J. Higgins: We appreciate Mr. Kazda's complimentary remark on the clarity of our exposition of Bode's method of return-voltage differences and our extension of it to include the hitherto untreated domain of limited stability.

We also appreciate Mr. Weiss's commendation. Contrary to a remark in his discussion, systems with distributed parameters are encompassed in this paper, as is specifically indicated in the beginning of the paper, where we state that although direct discussion is restricted to lumped-parameter systems, the analysis set out in this paper is "immediately possible by well-known theory." The only difference in the two types of systems is that whereas the transfer function of a lumped-parameter component is algebraic in nature, hence has a finite number of poles and of zeros, the transfer function of a distributed parameter network is transcendental in nature, hence may have an infinite number of poles, or of zeros, or of both. However, this is merely a matter of degree in the possible number of poles or zeros; the application of the theory in this paper to establish the nature of the stability of a distributed-parameter system proceeds in general in exactly the same manner as for lumped-parameter systems. Indeed, such fact is evidenced, e.g., in another manner in Dzung's paper,[1] where this author ascertains the stability of both lumped-parameter and distributed-parameter single-loop systems by his conformal mapping criterion (essentially, a reinterpretation of the significance of the Nyquist diagram plot) by the same mode of analysis.

However, systems with time-varying coefficients cannot (easily at least) be investigated by conjunction of transfer functions and Nyquist diagrams. Rather, investigation of stability must be effected by determination of the asymptotic nature of the solutions through the use of classical

differential equation theory, in the manner evidenced in Kirby's papers[2,3] or Bellman's recent excellent book,[4] to which attention is directed for an account of the corresponding mode of analysis. It is for such reason, therefore, that time-varying coefficient systems do not fall within the province of the present paper which deals with stability analysis pertinent to constant parameter systems.

We believe with Mr. Thaler that the present paper is the first in which return differences have been used for servomechanism and feedback control analysis. However, we would emphasize that whereas return-voltage difference theory in terms of complex-number impedances and admittances was developed by Dr. Bode for investigating nonlimited stability in feedback amplifiers and is extended in the present paper to encompass investigation of limited stability of feedback amplifiers as well, for the investigation of servomechanism and feedback control systems, the theory of return-signal differences in terms of Laplace-transform transfer functions, as developed in Section 6, is used. The two approaches are obviously somewhat analogous in the general course of procedure, but a significant difference exists between return-voltage difference analysis and return-signal difference analysis. Thus, to cite one essential difference other than as already noted regarding the respective use of complex-number admittances and Laplace-transform transfer functions: a return-voltage difference is a voltage quantity, but a return-signal difference may be an angle, velocity, current, or other quantity (though in a particular problem it may be a voltage quantity also).

Mr. Thaler's interesting, alternative derivation of $F_1(s)$ and $F_2(s)$ directly from the block circuit diagram of Fig. 7, rather than from the general determinantal (not "matrix," as he remarks) equation 28, ought prove of interest to numerous readers.

We appreciate Mr. Jones' opinion that significant educational advantages may stem from stability analysis conducted directly from the differential equations characterizing the system. However, use of this approach for analysis in general is directly contrary to current usage since most textbook and periodical literature[6] (at other than an elementary level) in all languages are written in terms of the Laplace-transform transfer-function approach because of the very considerable advantages that attend use of this approach. Moreover, the polynomial whose roots must be determined is essentially the same whether obtained from the Laplace-transform approach or the direct-differential-equation approach urged by Mr. Jones. The prime advantage attending Mr. Jones' suggestion is, of course, that the student using this approach need not learn transform theory. However, as for work or study at a consequential level in modern servomechanism or feedback control theory, the student must eventually gain a knowledge of basic transform theory, so he might as well master this at the outset of his study.

We agree with Mr. Jones that it is unfortunate that in the development of the use of the Nyquist diagram in control theory it early became customary to traverse the contour of the s-plane in the clockwise direction; whereas in complex variable theory

traverse in the counterclockwise direction is taken as traverse in the positive sense. However, this difference in direction of traverse is now so firmly established in the control literature that we believe that it can hardly be reversed. Our own approach in teaching is to point out the difference in the two areas, and then to proceed with the now well-established servomechanism convention of clockwise traverse.

We are grateful for Mr. Zilmer's excellent symmary of the essentials of our paper. We agree with Mr. Skiles' and Mr. Thaler's statements that the definitions of stability

advanced in most servomechanism books are rather loosely phrased. We concur with Mr. Skiles' statement that hitherto little has been published on the treatment of multiple-loop servomechanisms by the use of Nyquist diagrams on a per-loop basis. In conclusion, we extend sincere thanks to the discussers for their many interesting comments, illuminative remarks, and complimentary statements.

REFERENCES

1. AUTOMATIC AND MANUAL CONTROL. Butterworths Scientific Publications, London, England,

1952, "The Stability Criterion," L. S. Dzung, pp. 13–23.
2. STABILITY OF SERVOMECHANISMS WITH LINEARLY VARYING ELEMENTS, M. J. Kirby. AIEE Transactions, vol. 69, pt. II, 1950, pp. 1662–68.
3. STABILITY OF VARYING-ELEMENT SERVO-MECHANISMS WITH POLYNOMIAL COEFFICIENTS, M. J. Kirby, R. M. Giulianelli. AIEE Transactions, vol. 70, pt. II, 1951, pp. 1447–51.
4. STABILITY THEORY OF DIFFERENTIAL EQUATIONS (book), R. Bellman. McGraw-Hill Book Company, Inc., New York, N. Y., 1953.
5. BASIC BOOKS FOR YOUR CONTROL ENGINEERING LIBRARY: II—SERVOMECHANISMS, T. J. Higgins. Control Engineering, New York, N. Y., vol. 1, Dec. 1954, pp. 48–51. Lists 73 books.

Ignitron Multiple-Unit Cars for the New Haven Railroad

E. W. AMES
ASSOCIATE MEMBER AIEE

V. F. DOWDEN
ASSOCIATE MEMBER AIEF

ON APRIL 1, 1954, a 10-car train was placed in commuter service on the electrified lines of the New York, New Haven and Hartford Railroad, the first of 100 new multiple-unit cars. Modern in every respect, the outstanding feature is the use of ignitron rectifiers to convert alternating current collected from the trolley to direct current for operation of d-c traction motors. This is the first application in the world of this type of equipment in quantity. Its use follows naturally from the excellent results obtained with the ignitron rectifier type of motive power on a trial multiple-unit car in service on the Pennsylvania Railroad since 1949, followed by two locomotives which have been in operation on the same railroad since early 1952. These locomotives were nominally rated at 6,000 horsepower (hp), but have been operating at 8,400 hp each.

Service Operations

The new cars, built by the Pullman-Standard Car Manufacturing Company, weigh 78 tons without load and seat 120 passengers. Air conditioning and automatic heat control are incorporated to assure passenger comfort in every season. In contrast to existing motor cars, which employ hand acceleration and operate with trailers, the new cars are all equipped with four 100-hp motors and accelerate automatically at a fixed rate of 1.0 miles per hour per second, based on an average load of 60 passengers. All motor-car operation involves a minimum of terminal switching, assures uniform acceleration rates, and provides the greatest

protection of equipment and schedules in case of a power failure of one car in a train. The equipment is capable of operating in either local commuter service or in express service.

During the first 6 months the rectifier cars have operated a total of approximately 1,000,000 car miles. Counting days out of service for repairs and inspection, the availability factor has been 95 per cent. At present 50 trains are operated with the new cars on week days, 43 on Saturdays. On Sundays 33 trains are operated with the new cars, and in addition trains are at times operated from New Haven to New York on relief schedules. Two cars are assigned to trains operated on the New Canaan Branch and operate about 288 miles per day.

Seven of the new cars are combination cars, each having a passenger capacity of 92 and a baggage capacity of 10,000 pounds. Four of the cars are fitted for and used in club car service. One of these cars is equipped with a small buffet.

Performance

Fig. 1 illustrates the performance of a typical 6-car train operating from the 11,000-volt a-c trolley. Fig. 2 shows the performance from the 650-volt d-c third rail.

System of Control

Fig. 3 is a schematic diagram of the main circuit. It can be seen that the major portion of the main and auxiliary circuits are the same whether the car is operating in the a-c or the d-c zone. Only

the power supply differs. In the d-c zone the changeover switch connects the traction motor circuits and the d-c auxiliary circuits between the third rail shoes and ground. In the a-c zone the changeover switch connects these same traction motor and auxiliary circuits between the ignitron tube cathodes and the transformer mid-point through the main d-c reactor.

On the first notch of the master controllers, all motors and accelerating resistance are connected in series with the first step of field shunting to provide smooth starting. In both the a-c and d-c zones, resistor acceleration with a bridging transition is used with two steps of field shunting in the parallel combination for high-speed running. The acceleration is automatic at a fixed current rate under the control of the limit relay.

Rectifier D-C Supply

The d-c power in the a-c zone is supplied from a 25-cycle single-phase fullwave rectifier; refer to Fig. 3. The anode balance coils assure equal load division to the two rectifier tubes connected in parallel. The d-c chokes are used to limit the a-c component of the rectified direct current to values which the motors will commutate satisfactorily. The air brake compressor motor and car heaters do not require a d-c choke to operate from the rectified direct current; therefore they are connected from the cathode to the transformer mid-point.

Traction Motors

The traction motors are self-ventilated with single reduction gearing and flexible couplings. They are 325-volt d-c series-
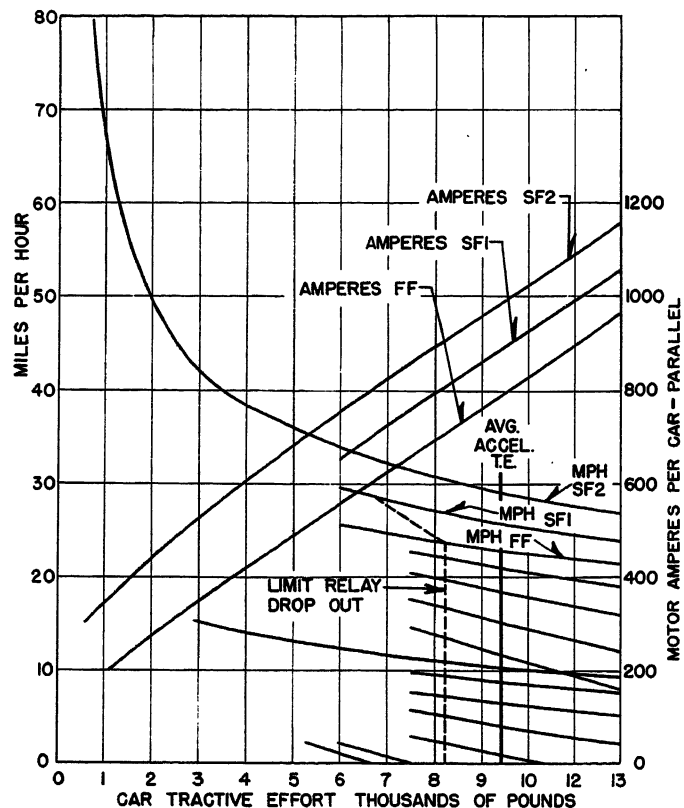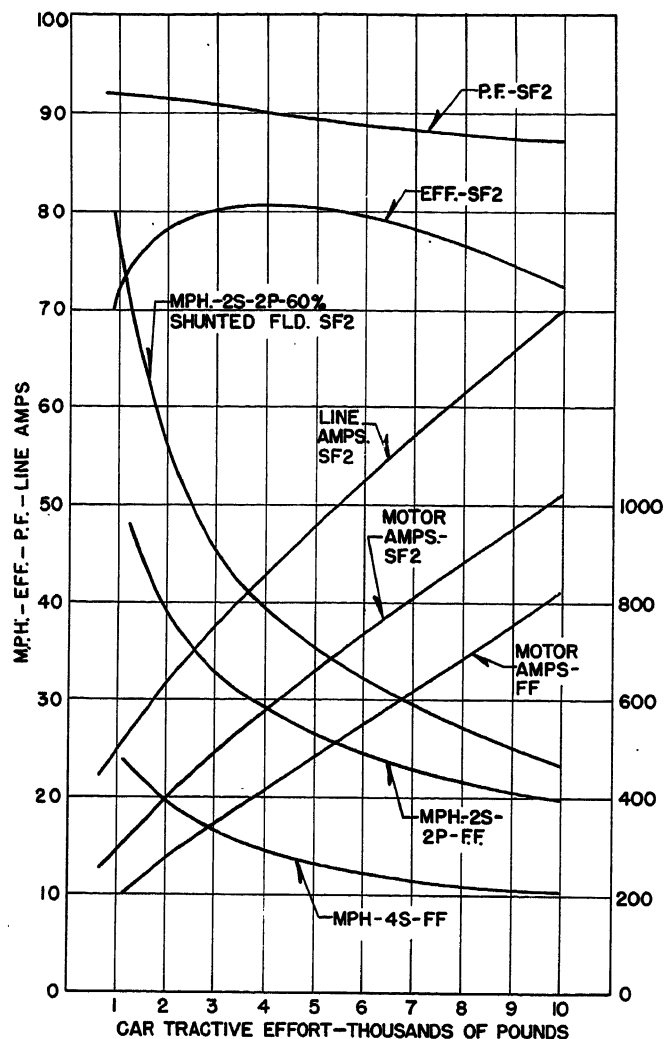
Fig. 1 (left). Performance of typical 6-car train for transmission systems and generator drops

Four WL.653-B ignitrons with four type-1451-A motors; 36-inch wheels, 122/23 gear ratio
Main line impedance=0.705+j1.69
Internal generated voltage=11,750
Average car heater+auxiliary motor load out of rectifier=50 amperes d-c
One can a-c filter
Distortion factors based on laboratory tests

Fig. 2 (above). Notching curve, third-rail zone 650 volts d-c

Four type-1451-A traction motors; 36-inch wheels; 122/23 gear ratio

wound commutating pole motors, and insulated for operation two in series on a maximum of 750 volts. The motors are designed for suspension from the truck transom, and are similar to those operating in rapid transit service in New York City.

Main Circuit Apparatus

The main transformer is an Inerteen-filled 380-kva transformer with an air-blast radiator, Inerteen to air, heat exchanger. One thousand cubic feet per minute of air is used for cooling. The transformer, designed for mounting under a standard American Association of Railroads center sill, is 52³/₄ inches wide, 29¹/₄ inches high, 94³/₈ inches long, and weighs 7,840 pounds. The primary winding is 11,000 volts, 25 cycles. The secondary voltage from the mid-point to either end is 780 volts with one tap at 226 volts from the transformer secondary mid-point for a-c auxiliary circuits. The a-c filter consists of two capacitors and resistors across the entire transformer secondary to suppress the induced currents in telephone lines adjacent to the

electrified lines in the a-c zone of operation.

The four WL-653B ignitron rectifier tubes are attached to a frame, insulated and shock mounted from the main box construction; see Fig. 4. Insulation to ground is provided by using 13 feet of rubber hose connected to the inlet and discharge headers of the tubes for the closed circuit cooling system. The water temperature is regulated by a temperature-sensing device and a 3-way valve to maintain the water temperature between 42 and 45 degrees centigrade (C) by passing water through the radiator for cooling or by-passing the radiator to retain heat in the water. Water heaters are turned on automatically whenever the water temperature is below 25 C [77 degrees Fahrenheit (F)]. The water system is protected from freezing by antifreeze. The radiator for the cooling system is located in the air intake to the blower. The air temperature is raised approximately 5 C and then utilized for cooling the main transformer and main d-c choke. There are two excitation circuits of the non-linear reactor type located at one end of the rectifier cubicle for excitation of the

rectifier tubes. Misfire lights, one of each tube, have been provided to enable maintenance personnel to check the rectifier operation. The rectifier cubicle has been constructed in the form of a letter T. It is 29 inches high, 65 inches long, and 49 inches at its widest part. Three felt-sealed, lift-off covers are provided for access to maintain the apparatus. The rectifier cubicle, located under the car, weighs about 1,880 pounds.

The major portion of the control apparatus is located in the main and auxiliary control box. The box contains the electropneumatic switches which serve as the a-c and d-c line switches, field shunting switches, the switches which short-circuit accelerating resistance steps, and those that set up the series and parallel motor combinations. Also placed in the box are the reverser, a-c/d-c changeover

Fig. 3. Schematic diagram

(Schematic wiring diagram of the ignitron multiple-unit car, including pantograph, lightning arrester, ground switch, thermal relay, PL relay, 1000-volt 25-cycle transformer, main d-c choke, limit relay, reverser, inductive field shunts, accelerating resistors R1–R4, ignitron tubes, anode balance coils, changeover switch, third rail shoes, motor generator set, air conditioning motors, car heaters, air brake compressor, and associated switches LS1–LS4.)

SEQUENCE CHART DC

CONT'R	STEP	LS-1	LS-2	LS-3	LS-4	JR	J	MG	R-1	R-2	R-3	R-4	F-1	F-2
OFF	1		O	O	O									O
	2		O	O	O									
2	3		O	O	O			O						
	4		O	O	O			O	O					
	5		O	O	O			O	O	O				
	6		O	O	O			O	O	O	O			
	7		O	O		O								
	TRAN		O	O		O	O							
	8		O	O				O						
	9		O	O				O	O					
3	10		O	O				O	O	O				
	11		O	O				O	O	O	O			
	12		O	O				O	O	O	O	O		
	13		O	O				O	O	O	O	O	O	
	14		O	O		O								O	O

SEQUENCE CHART AC

CONT'R	STEP	LS-1	LS-2	LS-3	LS-4	JR	J	MG	R-1	R-2	R-3	R-4	F-1	F-2
OFF	1	O	O	O	O	O								O
	2	O	O	O	O	O								
2	3	O	O	O	O	O		O						
	4	O	O	O	O	O		O	O					
	5	O	O	O	O	O		O	O	O				
	6	O	O	O	O	O		O	O	O	O			
	7	O	O	O	O		O							
	TRAN	O	O	O	O	O	O							
	8	O	O	O	O			O						
	9	O	O	O	O			O	O					
3	10	O	O	O	O			O	O	O				
	11	O	O	O	O			O	O	O	O			
	12	O	O	O	O			O	O	O	O	O		
	13	O	O	O	O			O	O	O	O	O	O	
		O	O	O	O		O						O	O

switch, operating and protective relays, auxiliary circuit breakers, electromagnetic contactors for car heaters, indicating lights, and a-c and d-c watt-hour meters. Fig. 5 shows the relative location of some of the apparatus in the main and auxiliary control box. The reverser and a-c/d-c changeover switch have handles extending through the bottom of the box so that they may be operated by hand if required.

The equipment boxes of previous multiple-unit cars operating on 650-volt third-rail systems have been mounted on insulators. The ignitron assembly, the main and auxiliary control box, and the regulator and control box for these new cars are bolted directly to the under frame of the cars without the use of insulators, to comply with Interstate Commerce Commission EX Parte 179. The additional insulation required in order to use a grounded box structure is pro-vided inside the boxes. Having these box structures grounded provides an added degree of safety for the personnel servicing the cars. Screens protect personnel from the ungrounded parts of the accelerating resistors, field shunts, auxiliary reactor, and a-c filter resistors. Fig. 6 shows a side view of the completed car. Figs. 7 and 8 show views of the car interior and operating position respectively.

Protective Apparatus

The main transformer is protected from primary overloads by the pantograph lowering relay. If the primary current exceeds 460 amperes, it is tripped magnetically, or if current in excess of the transformer rating continues over a predetermined time, it trips thermally. When the pantograph lowering relay trips it operates the pneumatic ground switch which grounds the line, causing the substation circuit breaker to trip. After the substation breaker trips, the pantograph is automatically lowered and locked down on the defective car. If the Inerteen of the main transformer exceeds 85 C, the traction motors are disconnected from the rectifier by opening LS3 and LS4; this lights an indicator.

If an arc back occurs, the LS1 and LS2 switches are opened by the a-c overload relay. The a-c overload can be reset from the master controller or by a push button which extends through the box cover. The d-c overload is a combination of two relays, an overload relay and a line switch relay. An overload in either traction motor circuit trips the d-c overload relay, which opens LS3 and LS4. Like the a-c overload relay, it may be reset from the master controller or manually by the pushbutton extending through the box cover.

The ground detector relay trips if a ground occurs on the secondary of the transformer, or any circuit connected to it, and lights an indicating light. The ground relay operates only in the a-c zone.

The rectifier is automatically shut down by thermostats if the tube jacket temperature exceeds 55 C (131 F) or if the water temperature falls below 5 C (41 F). If the water pressure, which is normally about 30 pounds, falls below 15 pounds an under-pressure switch shuts the rectifier down. Whenever any of these conditions are corrected, LS1 and
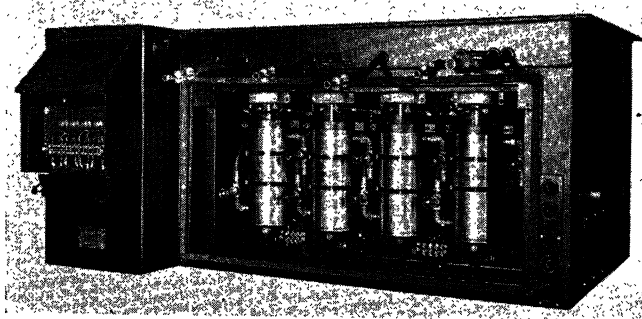
Fig. 4. Ignitron equipment box, rectifier tube compartment
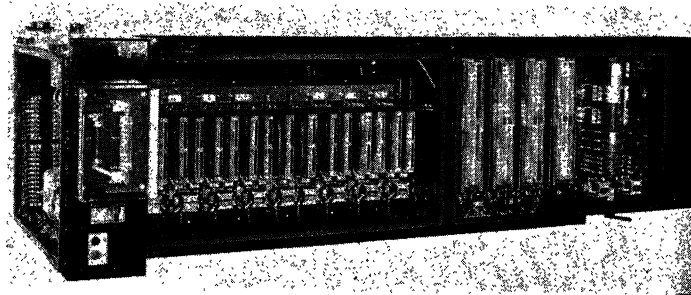
Fig. 5. Main and auxiliary control box, front view

LS2 are closed and the rectifier again supplies d-c power to the car.

A "power-on" relay is used to open the LS3 and LS4 switches of the main-motor circuits whenever there is a loss of d-c power due to third-rail gaps or bouncing pantograph. When LS3 and LS4 open, all main motor switches open. On reclosing LS3 and LS4, the control starts from the switching position and notches up automatically under the control of the limit relay.

Auxiliary Circuits

D-c motors fed from either rectified power or the third rail are used for all auxiliaries which are required in both the a-c and d-c zones. The motor-generator (m-g) set, air compressor, air-conditioning compressor, and condenser motors are 650-volt d-c motors. The car heaters are supplied from the rectified power or the third rail.

The transformer and reactor blower and the transformer Inerteen pump are required only for a-c operation. The motors are capacitor-start, capacitor-run induction motors which are fed from a 226-volt tap on the transformer secondary. The ignitron water heaters and anode heaters are fed from the third rail in the d-c zone and from the 226-volt tap through a 226 to 650-volt transformer in the a-c zone.

Sixty-four volts d-c for control, low-voltage auxiliaries, and battery charging is supplied from the generator of the m-g set. The motor of the m-g set is a 650-volt d-c compound motor which operates from either rectified or third rail direct current. The control for the m-g set is located in the regulator and control box.

The 64-volt control battery is made up of 32 lead-acid cells with a rating of 144 ampere-hours at the 2-hour rate. The entire battery is housed in a single box having a metal grating at the bottom and rear for drainage and ventilation. The battery floats on the generator with the charging voltage of 73 volts. No flushing

has been necessary between inspection periods. The battery provides a 1-hour emergency stand-by supply in case of power failure.

The car body and saloon are illuminated by fluorescent lamps operated from a 2.0-kw "Safety"-type motor-alternator. The lighting provides approximately 20 foot-candles at reading level. Twelve 60-volt 15-watt emergency lights are provided and are automatically controlled. Two 25-watt 60-volt incandescent flood lamps are provided in each vestibule. The marker lights are built in with 30-volt 30-watt lamps with series resistors. The headlight at each end is the conventional 60-volt 250-watt prefocused type with dimmer circuit.

The air-conditioning apparatus is rated at 8 tons dry. It is operated by a 2-speed motor direct-connected to the compressor. The evaporator is blown by a 60-volt 2.0-hp blower which distributes either heated or cooled air at a normal rate of 2,400 cubic feet per minute. When power fails the blower is automatically speeded up to deliver approximately 3,000 cubic feet per minute. This feature is necessary to provide increased ventilation, since during power failure the air conditioning cannot operate. The 650-volt d-c air-conditioning compressor and condenser motors operate in parallel on a single set of controls. The compressor motor is rated 11.5 hp and the condenser fan motor at 2.0 hp.

A total of 38.88 kw is provided for car

body heating. There are two vestibule blower heaters, each rated 1.5 kw, and one overhead heater unit rated at 21 kw arranged for three steps of control: 7 kw, 14 kw, and 21 kw. In addition, two circuits of wall heat, each 7.44 kw, and four body end blower heaters, each rated 750 watts, are provided. All heaters are thermostatically controlled except the two vestibule blower heaters which are manually controlled by the engineman. The wall heaters provide 60 F layover heat under the control of a separate thermostat.

Carbon Brush Life

One of the major limitations on the older classes of multiple-unit cars has always been the short life of carbon brushes, particularly those used in the traction motors. Experience to date indicates that brush life on the new cars may be anticipated as follows:

Traction motors, 85,000 to 90,000 miles
Ignitron pump motor, 3 years
M-g set motor, 9 months
M-g set generator, $4^{1}/_{2}$ months
Motor-alternator motor, 12 months
Motor-alternator generator, 3 years
Air-conditioning compressor motor, 2 years
Air-conditioning condenser motor, 3 years
Air brake compressor motor, 2 years

This anticipated brush life is based upon observations of performance during the first 6 months of operation. It is the opinion of the authors that the brush life can be extended in some cases either
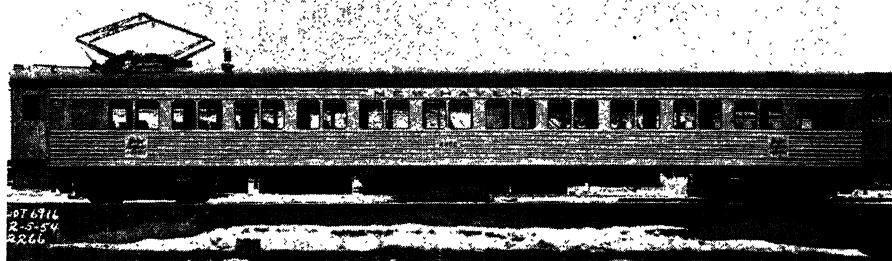
Fig. 6. Side view of completed car
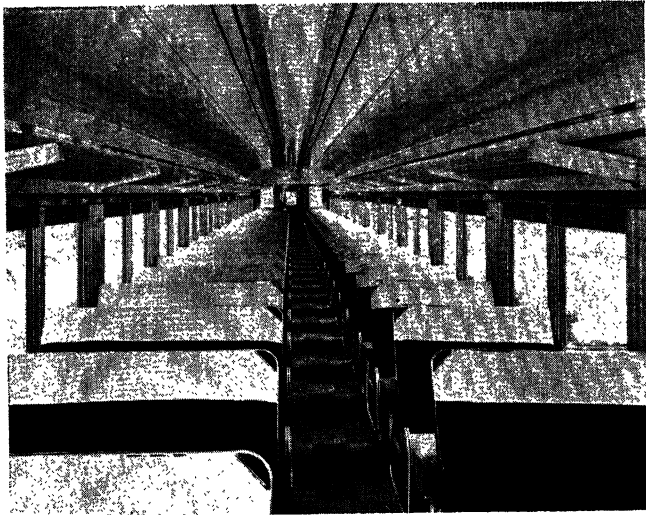
Fig. 7 (left). Interior view of car
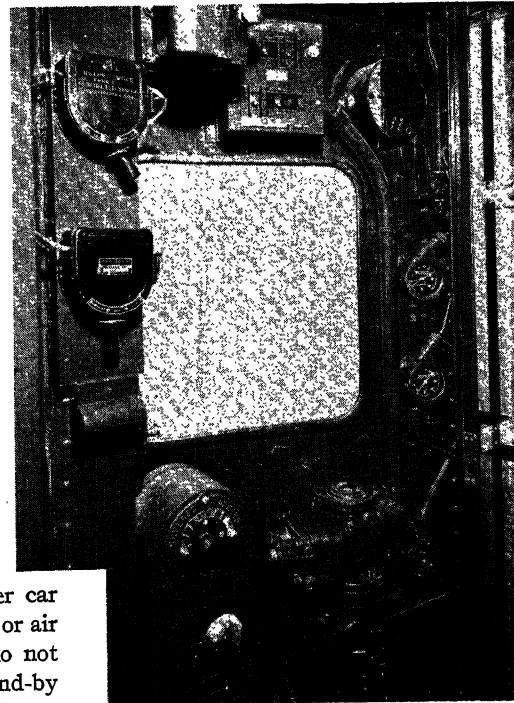
Fig. 8 (right). Car operating position

by change of brush grade or by adjustment of pressure.

Inspection Periods

The first inspection period on these cars was made after 4,000 miles of operation. Experience gained from this inspection led to making the second inspection after an additional 16,000 miles of operation. Certain other inspection restrictions may be imposed; however, experience so far indicates that the new equipment should require a minimum of maintenance.

Power Consumption

Preliminary tests on power consumption in local service show approximately 6.0 kw-hours per car mile, in express serv-

ice approximately 3.4 kw-hours per car mile. These figures include heating or air conditioning and auxiliaries, but do not include stand-by losses. A-c stand-by losses are estimated at about 19 kw-hours per car per hour, but operating conditions and practices have kept these losses to a minimum.

Conclusion

The use of the rectifier system is advantageous in that it combines the desirable characteristics of high-voltage a-c transmission with straight d-c traction motors and control equipment. Furthermore, auxiliary motors, such as the air compressor, the air-conditioning compressor, and condenser motors, are 650-volt d-c motors. The selection of 650-volt d-c motors for these auxiliaries made it pos-

sible to reduce the size of the control battery and the m-g set. It is estimated that the weight of the car would have been increased approximately 3,000 pounds if 64-volt air-conditioning motors and a larger battery had been used.

The rectifier-type motive power with its high power factor and efficiency, with the use of standard 650-volt d-c motors, is well suited to a-c electrification. The New Haven Railroad has an additional advantage by the use of rectifier motive power, since the cars are required to operate over 11 miles of the New York Central Railroad's 650-volt d-c third-rail system into Grand Central Terminal.

Discussion

H. F. Brown (Consulting Engineer, Westinghouse Electric International Corporation, New Haven, Conn.): The large number of cars of the rectifier type described in this paper and now in service changes the status of this type of equipment from an interesting experiment into a commercial reality. Certainly the mileage already run off of over 1,720,000 performed by 50 trains daily during the past 6 months is an indication of successful performance.

The authors did not specifically mention it, but those who have examined the commutators of the traction motors on these cars have noted their excellent appearance and condition even after mileages exceeding 75,000. Compared with the a-c/d-c series motors used on the motor cars which have been in service on this road for many years, there is every reason to believe that these new motors will have much lower maintenance expense.

Mention is made under "Power Consumption" of stand-by losses. Could the

authors amplify this term and state what these losses consist of? How do these losses compare with the older-type cars? Also, how does the power consumption in service compare with the older-type cars?

L. J. Hibbard (Westinghouse Electric Corporation, East Pittsburgh, Pa.): The New York, New Haven and Hartford Railroad Company, the Pennsylvania Railroad, and Westinghouse Electric Corporation pioneered the use of the high-voltage single-phase electrification system in this country. It is therefore fitting that the New Haven Railroad should be the one to make the first large-scale application of rectifier motive power in the world and that the Pennsylvania Railroad and Westinghouse should be the first to demonstrate the practicability of rectifier motive power.

An act by the New York State Legislature dated May 7, 1903, authorized the New York Central and the New York, New Haven and Hartford Railroads to run their trains through the Park Avenue Tunnel by electricity or by compressed air

Table I. Stand-by Kva at 11-Kv Trolley for Auxiliaries of Rectifier Car

	Kva
Transformer and rectifier blower = 19.6 × 226 =	4.43
Transformer Inerteen pump motor = 4.0 × 226 =	0.90
Rectifier firing circuits and anode heaters = 10 × 226 =	2.26
Main transformer (380 kva) excitation kva = 0.0332 × 380 =	12.60
Air-brake compressor motor (1/3 time) = (7 × 780)/3 =	1.82
Air-conditioning* compressor motor = 19 × 780 =	14.83
Air-conditioning* condensor fan motor = 2.35 × 780 =	1.82
Car heaters* at maximum heat = (38.88 × 780)/600 =	50.50
M-g† set = 26 × 780 =	20.50

* Either car heating or air conditioning is used depending on the weather.
† The m-g set supplies the following at 73 volts: Battery charging (144 AH at 2-hour rate). Ignitron water circulating pump 1 hp. Car body air-circulating fan 2 hp. Headlight 250 watts. Four vestibule lights, total 100 watts. Motor alternator 2-kw output supplies fluorescent lighting and water cooler.

or by any motive power other than steam which did not involve combustion in the motive units themselves. This act required that the change of motive power be made on or before July 1, 1908, and provided that a penalty of $500 per day be exacted on and after that date for failure to comply with its terms.

Prior to this, the Westinghouse Electric and Manufacturing Company had developed the high-tension single-phase system. The New York Central plans called for a relatively short electrified zone and they decided to adopt the 650-volt d-c third-rail system. The New Haven Railroad planned on an initial electrification to Stamford and New Haven with possible future extensions eastward and decided to adopt the high-voltage single-phase system. These three companies were also partners in a pioneering rectifier motive power application in 1913, 1914, and 1915.

A Pennsylvania Railroad combination baggage and passenger car (car no. 4692) was equipped with four type-*308* 225-hp 600-volt traction motors and d-c control equipment borrowed from the Long Island Railroad and with transformers, multianode rectifiers, and accessory a-c control equipment supplied by Westinghouse. This car was tested on the Westinghouse test track at East Pittsburgh and then placed in revenue service on the New York, New Haven and Hartford Railroad. It hauled two trail cars and performed 8,757 miles in revenue service on the Harlem River branch and 13,588 miles on the New Canaan branch for a total revenue service of 22,345 miles.

The Pennsylvania Railroad and Westinghouse again pioneered rectifier car no. 4561 in 1949. Here again, the Pennsylvania Railroad supplied a combination passenger and baggage car with Long Island Railroad d-c motor and control equipment (two 559 *DR3* motors and associated control) while Westinghouse supplied the transformer and rectifier equipment. This car started operation July 14, 1949, and has been in continuous service since then.

The Pennsylvania then purchased two 2-cab rectifier locomotives which have been in revenue service for approximately 3 years. The New Haven Railroad then purchased 100 new rectifier multiple-unit cars and ten rectifier locomotives.

E. W. Ames and V. F. Dowden: Table I gives the kilovolt-amperes (kva) required for all auxiliaries of the rectifier car from test readings of amperes for each circuit. Table II is tabulated on the same basis as Table I for the stand-by kva at 11-kv

Table II. Stand-by Kva at 11-Kv Trolley for 409 Motor Car and Two Trail Cars

	Kva
Transformer and traction motor blower $37 \times 420 =$	15.6
Main transformer (805 kva) excitation kva $3.4 \times 1,100 =$	37.4
Air-brake compressor motor (1/3 time) $(36 \times 347)/3 =$	4.2
Car heaters* including two trail cars =	150.0
M-g set 1-kw output	

* Car heating is maximum available.

trolley for the 409 motor car and two trail cars. The rectifier cars accelerate at a 1-mile-per-hour-per-second rate while the older cars accelerate between the 0.5- and 0.75-mile-per-hour-per-second rate.

The new cars, with a higher acceleration rate, have enabled the railroad to reduce the scheduled time for a run. The new cars offer many passenger comforts, such as automatically controlled heating or air conditioning, fluorescent lighting, forced-air circulation, and many other comforts not available on the older-type equipment. Power consumption for the 409 motor car with two trailers is not available on the same basis as that stated in the paper.

Conditional Feedback Systems—A New Approach to Feedback Control

G. LANG
NONMEMBER AIEE

J. M. HAM
MEMBER AIEE

Synopsis: In the classical single-loop feedback system, feedback acts not only to modify the influence of disturbances but also to determine the basic character of the input-output response. The inherently close association of these two effects has been a constant trial for designers and students since the inception of classical feedback, and has usually required that design requirements be compromised. Basically new configurations for feedback systems are introduced in which these effects of feedback are separated. In the new systems, feedback acts solely to reduce the influence of disturbances and thus to determine the response of the system to external loads and internal parameter variations. The character of the input-output response is independent of transmission around the feedback loop. The term "conditional feedback" has been introduced to distinguish the limited role that feedback plays in these new systems as compared with the classical system. Conditional feedback systems permit requirements on input-output response and on disturbance-output response to be met independently and offer a broad new range of performance characteristics for both linear and nonlinear systems in which there is substantial energy storage.

THE classes of closed-loop systems introduced in this paper were evolved after an unsophisticated review of the problems encountered and the solutions accepted in present-day engineering practice in control and regulation. In the design of these new systems, the need for compromise in performance specifications as a result of conflicts between input-output response and disturbance-output response has largely been eliminated. Such conflicts are inherent to the conventional feedback control system because of the intimate relationship existing between the input-output transmission characteristic and the loop transmission characteristic. Since the latter is dictated by requirements for loop stability and for the suppression of disturbances, little variation is allowed the former, and at that not without considerable compromise. A useful analogy for the foregoing condition is found in the intimate connection that exists between the frequency response and the phase characteristic of minimum phase networks. These restrictions can be overcome by the use of networks having nonminimum phase characteristics. In a like manner, a basic change in topology can be used to extend the present boundaries of feedback control systems.[1-4]

Classification of Control Problems

The classification of control problems adhered to in this paper distinguishes two basic categories defined as follows:

1. The servo class: A servo problem requires the generation of an output signal that bears a prescribed functional relationship to an input signal. This problem is commonly met in all types of signal transmission. A servo system is applied to the solution of a servo problem.

2. The regulator class: A regulator problem requires the elimination or reduction of the effects on a controlled quantity of extraneous and generally poorly defined disturbances. A regulator system is applied to the solution of a regulator problem.

Fig. 1 (left). A representative feedback system

Fig. 2 (right). Configuration of a conditional feedback system

Feedback control systems are applied to both the aforementioned problems because they can act simultaneously as a servo and as a regulator. This property has been found to be of the utmost value in the solution of the many problems in which both servo and regulator requirements exist. Purely servo problems do not necessarily require feedback for their solution, nor do purely regulator problems, e.g., the glow-discharge tube is used to provide voltage regulation without feedback.

Conventional Feedback Control Systems

The dual role played by the conventional feedback control system is easily exposed by considering the Laplace transform C of the output signal c that results when an additive signal u_n having the transform U_n is applied successively to the nodes $n = 1, 2, 3, 4$, and 5, in the representative feedback system of Fig. 1. For an input signal r with the transform R and successive additive signals u_n, the successive outputs c_n have the transforms

$$C_1 = \frac{G_1 G_2 U_1}{1 + HG_1 G_2} + \frac{G_1 G_2 R}{1 + HG_1 G_2} = F_1 U_1 + FR$$

$$C_2 = \frac{G_1 U_2}{1 + HG_1 G_2} + \frac{G_1 G_2 R}{1 + HG_1 G_2} = F_2 U_2 + FR$$

$$C_3 = \frac{U_3}{1 + HG_1 G_2} + \frac{G_1 G_2 R}{1 + HG_1 G_2} = F_3 U_3 + FR \quad (1)$$

$$C_4 = \frac{-G_1 G_2 U_4 H}{1 + HG_1 G_2} + \frac{G_1 G_2 R}{1 + HG_1 G_2} = F_4 U_4 + FR$$

$$C_5 = \frac{-G_1 G_2 U_5}{1 + HG_1 G_2} + \frac{G_1 G_2 R}{1 + HG_1 G_2} = F_5 U_5 + FR$$

The response terms on the right of equation 1 express the principle of superposition for the signals u_n and r. It follows from equation 1 that

$$F_1 = F$$

$$F_2 = \frac{F}{G_2}$$

$$F_3 = \frac{F}{G_1 G_2} \quad (2)$$

$$F_4 = -HF$$

$$F_5 = -F$$

From equation 2 it is clear that disturbances at nodes 2 and 3 are quelled by concentrating gain in G_2 and G_1, whereas disturbances at the input to H (considered as an output-sensing device) are of a particularly troublesome nature and cannot be reduced without lowering the gain or bandwidth of H. This latter view points out the necessity for care in the choice of components for output sensing.

For a linear system, a disturbance at node 2 can be considered as a disturbance of different amplitude and frequency distribution entering at node 3. Herein all external disturbances will be represented in terms of an equivalent load disturbance u.

Disturbances

The importance of knowing the disturbance problems associated with a particular system design cannot be too heavily stressed. If there are no external disturbances and available system components are linear and not subject to parameter variations, an open-loop system is ideally suited to most servo problems provided a suitable open-loop transfer function can be obtained.

Open-loop systems have the particular advantage that real time delays are often unimportant. It is the character of the delayed response that is usually of major interest. Further, in such systems, adequate control of the influence of certain nonlinearities and load disturbances can be gained by providing stable compensating nonlinear elements and an output with a suitably low driving-point impedance. The influence of disturbances entering the system between input and output can often be reduced by isolating the system from known sources of disturbance, e.g., transmission lines may be transposed to reduce the effect of induced signals.

There will remain, however, many classes of parameter variation, nonlinearity, and internal disturbance that are not amenable to treatment by such techniques. For these a closed-loop regulating system is required. A closed loop will modify the influence of those system variations for which a closed loop is not necessary, but improved system

performance will usually be obtained when all variations susceptible to open-loop compensation are so treated. Closed-loop regulation is required in nearly all machine- and process-control problems. Another important role for closed-loop control is to modify or synthesize transfer functions which, in many cases, would be most intractable to synthesis in a simple manner by open-loop techniques.[5]

Configuration and Basic Properties of Conditional Feedback Systems

Two distinct control problems have been identified: the servo or signal transmission problem and the regulator or disturbance suppression problem. The performance requirements for signal transmission and disturbance suppression in practical applications are usually distinct, but the signal and disturbance behavior characteristics of classical feedback systems have been shown to be inherently interdependent. Hence, the design requirements for input-output response and disturbance-output response had to be compromised. Since feedback in many control problems is essential solely to effect a suitable reduction in the influence of disturbances, the question arises whether there are systems in which the use of feedback can be so restricted. Systems possessing this property are called conditional feedback systems. A basic configuration for a linear conditional feedback system is shown in Fig. 2. In Fig. 2, G_1 is the Laplace transfer function describing the main transducer, r is the input signal, u is a disturbance, and c is the output. For the system shown, the transform C of the output c is

$$C = \frac{1}{1 + G_1 G_2 H} U + \frac{AG_1\left(1 + \frac{B}{A} G_2\right) R}{1 + G_1 G_2 H} \quad (3)$$

In equation 3, let B be defined by the equality

$$G_2 \frac{B}{A} = G_1 H G_2$$

or

$$B = AG_1 H \quad (4)$$

Then equation 3 becomes

Lang, Ham—Conditional Feedback Systems
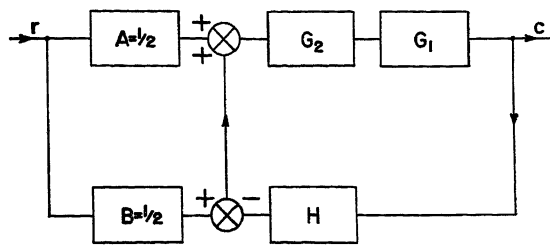
Fig. 3 (left). The conditional configuration equivalent to the classical system

Fig. 4 (right). Equivalence of conditional system to a classical system with a prefilter
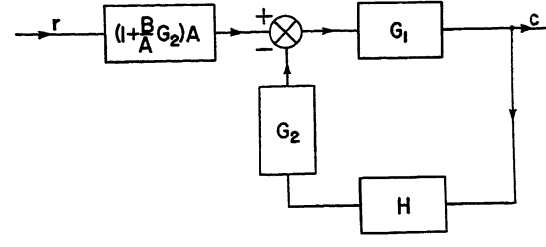
$$C = \frac{1}{1+G_1G_2H}U + AG_1R \qquad (5)$$

from which

$$\left(\frac{C}{R}\right)_{u=0} = AG_1 \qquad (6)$$

and

$$\left(\frac{C}{U}\right)_{r=0} = \frac{1}{1+G_1G_2H} \qquad (7)$$

Equation 6 shows that when B is defined by equation 4 the input-output response of the system is unaffected by the feedback loop and can be of a basically different character from the disturbance-output response in equation 7, the form of which is completely determined by the feedback loop. The significance of these equations is immense for they show that a conditional feedback configuration permits design requirements on input-output response and on disturbance-output response to be considered independently. This fact implies that a broad new range of performance characteristics is available from conditional feedback systems.

Before giving a design procedure and illustrative design for the conditional system of Fig. 2, some discussion of its internal behavior is in order. If the transfer function B satisfies equation 4, the feedback signal b in the absence of disturbances is exactly equal to the signal v produced at the output of B by the input signal r. Hence the signal e at the output of the lower comparator is identically zero and there is no feedback. The transfer ratio C/R for the signal transmission is then simply the product of the transfer functions in the direct path from input to output; see upper branch of the block diagram in Fig. 2.

These facts point to a special significance for the system component described by the transfer function B. The function of B is most readily suggested by considering the system when $H=1$. In this case, the feedback signal $b=c$. Hence when B is defined by equation 4, the output signal from B, namely v, is equal to the output signal c. Clearly then B represents the desired input-output transfer function and the signal v represents the desired output response. For this reason, the system component described

by the transfer function B in Fig. 2 will be called a "reference model." The reference model is an undisturbed representation of the transfer characteristics of the main transducer as defined in equation 4, and will usually be realized with simple R, L, and C elements. When H is other than unity, v is the desired output signal as modified by H. H commonly describes the output-sensing device.

An interesting comparison between the classical feedback system of Fig. 1 and the conditional system of Fig. 2 is obtained by developing a conditional configuration that is equivalent to the classical system. This system is shown in Fig. 3. From equation 3 it follows that, if $A=B=1/2$, the system in Fig. 3 behaves as the classical system of Fig. 1. Clearly the classical system calls for an ideal reference model. Since ideal behavior is not to be expected from practical equipment, it is not surprising that the use of an ideal reference model entails special system performance limitations.

W. K. Linvill pointed out to the authors that the conditional feedback system of Fig. 2 is equivalent to a classical system with a prefilter, as shown in Fig. 4. In Fig. 4, the response C to R and U is that given by equation 3. When the transfer function B in Fig. 4 is defined by equation 4, the prefilter has the particular transfer function A $(1+G_1G_2H)$, the response ratio $C/R=AG_1$, and there is no signal in the feedback path; hence, the combination of this particular prefilter and the classical system has the properties of a conditional system. It should be observed that both of the foregoing equivalences are valid only when the system components are linear.

Since feedback in a conditional feedback system is used solely to reduce the influence of disturbances, it is important to examine carefully the action of disturbances on this new configuration. In a conditional feedback system such as that of Fig. 2, there is no loop feedback if there are no disturbances. The error signal e obtained by comparing the desired signal v with the fed back signal b serves as a measure for the effect of disturbances. In developing conditional systems, it has been observed that in the comparison of v with b either a subtraction or a ratio

operation may be used. If a subtractive comparison is used, an additive correction is made at the input to G_1, as in Fig. 2. If a ratio comparison is used, the signal input to G_1 is corrected by multiplying the signal output from A, namely m, by some function of the signal ϵ from the ratio comparator.

The action of two types of disturbances will be examined. First consider the steady-state response of the system in Fig. 2 to a constant additive disturbance u. The output signal c will be in error by a constant and the error signal ϵ will be constant whether or not there is an input signal v. Next consider the system of Fig. 2 with the subtractive comparator replaced by a ratio comparator and the additive corrector replaced by a multiplicative corrector. Let there be a constant multiplicative disturbance such as may be caused by a decrease in the static gain factor of G_1. The output signal c will be in error by a constant factor and the error signal ϵ will be constant. However, if additive and multiplicative disturbances occur together the error signal ϵ at the output of both a subtractive and a ratio comparator will contain frequency components of the input signal. These observations suggest that the nature of disturbances has an important bearing on the design of conditional feedback systems; this fact is discussed later in the paper.

The response of the linear conditional feedback system in Fig. 2 to an additive disturbance of general form is described by equation 7. The form of this response is controlled by an appropriate choice for the transfer function G_2. All of conventional feedback theory and design technique is relevant to making this selection.

It is important to observe that disturbances influencing the system component, described by A in Fig. 1, without affecting the reference model B, are compensated by feedback through G_2 to the output of A. Parameter variations in the components of a conditional feedback system such as that shown in Fig. 2 have essentially the same influence on the output as the corresponding variations in the classical feedback system of Fig. 1. Hence most of the literature on parameter sensitives is relevant. To illustrate this fact

Lang, Ham—Conditional Feedback Systems

consider a variation dG_1 in the transfer function of the main transducer. The corresponding variation in the transfer function of the output signal is dC. For the classical system of Fig. 1

$$dC = \frac{1}{1+G_1G_2H} \frac{C}{G_1} dG_1 \qquad (8)$$

For the conditional system, the variation dC has exactly the same form.

The foregoing discussion of a linear conditional feedback system has shown that this new configuration permits the independent control of the input-output response and of the disturbance-output response. The disturbance-output response can always be made as good as that of the corresponding classical system while the input-output response can have forms hitherto unrealizable. A design procedure for conditional feedback systems based on the foregoing development is described in the following.

Design Procedure and Illustrative Example

The special properties of conditional feedback systems as developed in the foregoing lead to a design procedure that is unusually direct and simple. A suggested procedure is outlined in the following. The particular significance of the steps will be illustrated in a sample design.

1. From a careful study of the system problem determine the required input-output response characteristics and the required disturbance-output response characteristics; special care should be given in determining the basic nature of disturbances.

2. Select a main transducer and associate with it compensating elements to realize the desired input-output response. This step involves the choice of the transfer functions G_1 and A in Fig. 2.

3. Select a suitable output-sensing device and determine its transfer function H.

4. Design an undisturbed reference model described by the transfer function $B = AG_1H$.

5. Design a loop-compensating network described by the transfer function G_2 to realize the desired disturbance-output response to the extent that the fundamental requirement of loop stability allows.

These steps in design form a direct sequence in which there is no need for compromise among the steps. To illustrate the design procedure for conditional feedback systems a sample design will be given for a conditional system and a corresponding classical system that employs the same main transducer. The feedback loop transmission function is to

be the same in both systems so that a direct comparison of performance characteristics is justified.

EXAMPLE

It is desired to construct a positional servomechanism using a main transducer described by the transfer function

$$G_1 = \frac{K_1 \epsilon^{-T_1 s}}{s}; \quad K_1 = 10, \quad T_1 = 0.2 \text{ second} \qquad (9)$$

G_1 corresponds to a time-delayed integration. Pure time delays in the loop of a classical feedback system impose a definite limit on the bandwidth that can be realized in input-output response. Such is not the case when a conditional feedback system is employed. The configuration of the conditional system is shown in Fig. 2. The corresponding classical system is shown in Fig. 1. The steps in designing the conditional system are now given.

From equation 6 it follows that the transfer function A must be synthesized to make the product AG_1 represent the desired form of the input-output response. Suppose that the desired response to a unit step has the form

$$c = \left(1 - \epsilon^{\frac{-(t-T_1)}{T}}\right); \quad T_1 = 0.2,$$
$$T = 0.01 \text{ second} \qquad (10)$$

From equation 10 the input-output response transform is

$$\frac{C}{R} = \frac{\epsilon^{-T_1 s}}{1+Ts} \qquad (11)$$

From equation 11 it follows that a suitable form for A is

$$A = \frac{K_0 T_0 s}{1+T_0 s} \qquad (12)$$

From equations 9 and 11, the values of K_0 and T_0 are given as

$$T_0 = T = 0.01,$$

$$K_0 = \frac{1}{T_0 K_1} = 10 \qquad (13)$$

The transfer function A is readily realized with gain and a simple resistance-capacitance network.

It is important to note that it is not necessary to effect the complete compensation of the main transducer with elements placed in the box marked A. The synthesis of the desired input-output response transform may include tandem compensation of G_1 placed in the upper branch of the feedback loop, classical feedback around G_1, and the like. However, if the feedback loop is required to have nonzero transmission at zero frequency to fulfill its regulating function,

compensating elements which do not transmit at zero frequency must be placed in A.

Now an output-sensing device may be selected and its transfer function H determined. To simplify the details of this design example, H is considered to be unity. Design of the undisturbed reference model is now carried out. From equation 4 and the condition $H=1$, it follows that the transfer function B for this model is

$$B = \frac{C}{R} = \frac{\epsilon^{-T_1 s}}{1+Ts}; \quad T_1 = 0.2,$$
$$T = 0.01 \text{ second} \qquad (14)$$

If the time delay T_1 of the main transducer is subject to parameter variation, the value used in equation 14 for designing the reference model is some mean value. In practice B will be realized with miniature elements having tolerances on their characteristics established by the allowable tolerances in the response ratio C/R.

To complete the design of the conditional feedback system, a loop-compensating transfer function G_2 has to be designed to realize the desired disturbance-output response, as defined by equation 7. Since it is desired to compare the performance of the conditional system with that of the corresponding classical system, the selection of G_2 will be based on the following considerations. The classical system in Fig. 1 corresponds to the conditional system in Fig. 2 when the function G_1, G_2, and H are the same. The disturbance-output response of these systems is identical and is described by equation 7. Since the input-output response of the conditional system is independent of G_2, the classical system will compare most favorably with the conditional system if the form of G_2 is selected to obtain the best input-output response from the classical system.

The response ratio C/R for the classical system in Fig. 1 is

$$\frac{C}{R} = \frac{G_1 G_2}{1+HG_1 G_2}; \quad H=1 \qquad (15)$$

In equation 15, G_2 is to be designed to obtain the best response ratio. The classical cut-and-try procedure is followed. The basic form for G_2 is taken to be

$$G_2 = K_2 \frac{\frac{1}{a_d}T_d s+1}{T_d s+1} \frac{T_i s+1}{\frac{1}{a_i}T_i s+1} \qquad (16)$$

G_2 can be realized with gain, a lead network, and a lag network. The parameters K_2, a_d, T_d, a_i, and T_i are to be selected.

Suppose that a static gain factor

Lang, Ham—Conditional Feedback Systems

Fig. 5. Step responses of the conditional and classical design

Fig. 7. A simplified multiplicative system
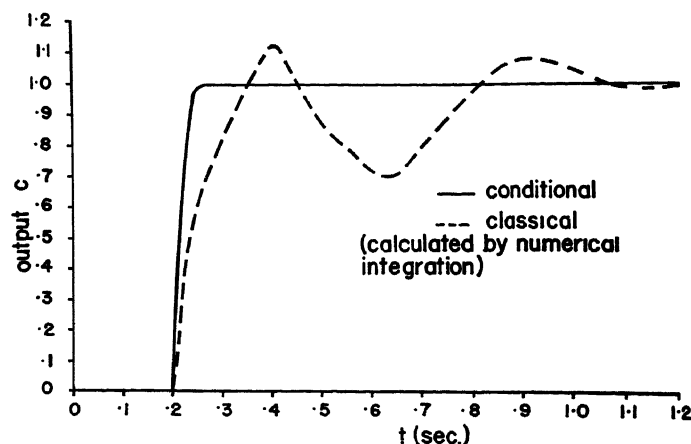
$$K = \lim_{s \to 0} sHG_1G_2 = 100 \qquad (17)$$

is required in the feedback loop. From equations 9 and 17 it follows that

$$K = K_1K_3$$

or

$$K_3 = \frac{K}{K_1} = 10 \qquad (18)$$

In equation 16, the gain factor a_d of the lead network is arbitrarily assigned the value 0.1. From a polar plot of G_1, the value $T_d = 0.025$ is selected for the time constant of the lead network. On the assumption that the lag network acts as a gain factor a_t at frequencies in the neighborhood of the critical point in the G_1G_2 plane, a_t is assigned a value $a_t = 1/40$, which results in a response peak of about 1.3 in ratio C/R. The time constant T_t of the lag network is assigned a value $T_t = 2.5$ seconds, which results in a phase shift through the lag network of less than 2 degrees at frequencies in the neighborhood of the critical point in the

G_1G_2 plane. The open-loop response of the classical system is thus defined as

$$G_1G_2 = \frac{100e^{-0.2s}}{s} \frac{0.25s+1}{0.025s+1} \frac{2.5s+1}{100s+1} \qquad (19)$$

The input-output response of the conditional system as defined by equation 11 is now to be compared with the input-output response of the corresponding classical system as defined by equations 9, 19, and 15. In Fig. 5, the responses of the two systems to a unit step are compared and in Fig. 6 the magnitude and phase characteristics of the response ratios C/R are compared.

In evaluating the response comparisons in Figs. 5 and 6, it is important to remember that the disturbance-output responses of the two systems are identical. The conditional system clearly has much superior input-output response characteristics. It is important again to emphasize that there is freedom of choice for the character of the input-output response in a conditional feedback system.
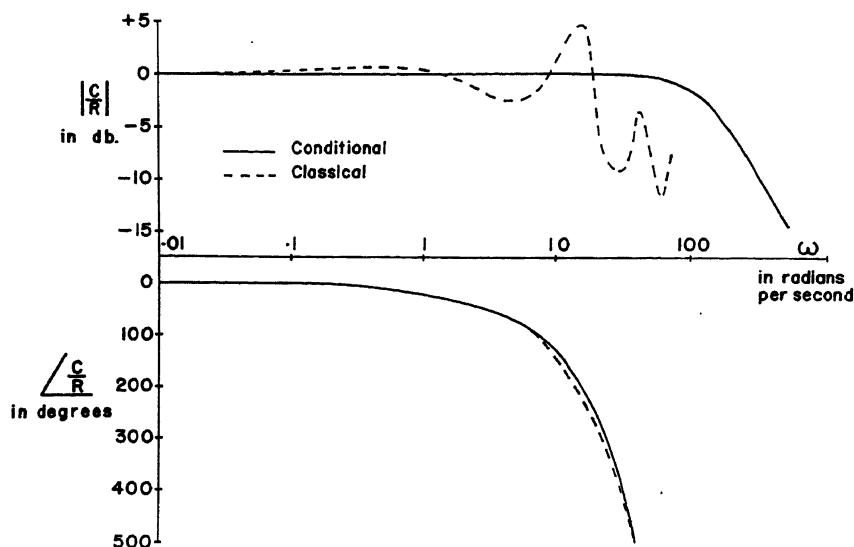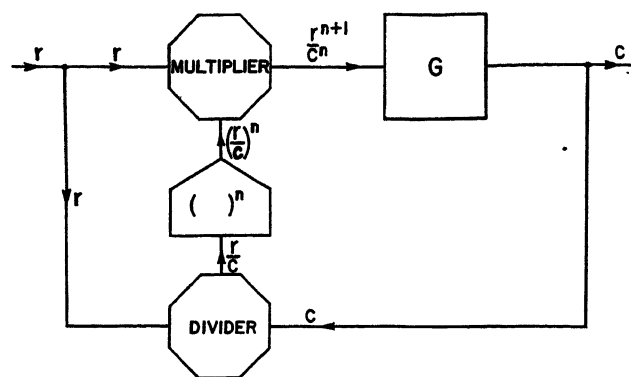
In a classical system, the effort to realize maximum bandwidth in the input-output response almost always leads to underdamped oscillatory characteristics. This fact is particularly true when pure time delays are present in the feedback loop.

The foregoing example is suggestive of the extended range of performance characteristic available to feedback system designers* when conditional configurations are used. It is also important to note that the input-output bandwidth may be reduced while the disturbance output bandwidth is maintained.

Apart from having basic significance in the domain of linear systems, conditional feedback configurations have certain special merits for systems containing nonlinearities and, indeed, a new class of nonlinear servomechanism has been evolved. The extension of the concept of conditional feedback to nonlinear systems is discussed in the following.

Extension of the Principle to Nonlinear Systems

DISTURBANCES

The casual observation that an error could be properly measured as a ratio rather than as a difference has led to a rather novel embodiment of the conditional topology and to some interesting conclusions concerning the nature of disturbances. It has been remarked previously that a steady-state multiplicative disturbance results in a constant error signal when the error signal is measured as the ratio of the undisturbed signal to the disturbed signal in a conditional system. In a like manner, a steady-state additive disturbance will result in a constant error signal when the error is measured as a difference. Hence, it is useful to classify arbitrary disturbances as being dominantly multiplicative or dominantly additive in character accordingly as the

Fig. 6. Frequency responses of the conditional and classical designs
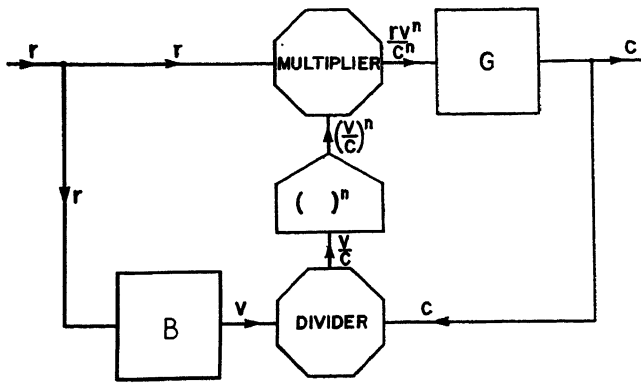
Fig. 8 (left). A multiplicative conditional servo

Fig. 9 (right). The incremental response of a simple multiplicative servo

o - calculated points
x - points measured in experimental servo.
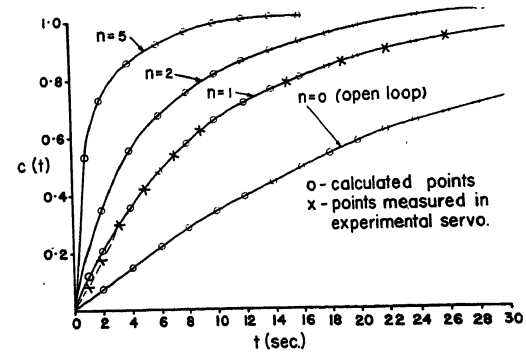
ratio measured error or the difference measured error has the lower bandwidth. A significant difference in bandwidth in these measurements, when interpreted in terms of basic bandwidth restrictions on the feedback loop, will suggest the adoption of a particular type of conditional system for dealing with a disturbance.

In view of these observations, conditional feedback systems have been classified as being either multiplicative or additive in character. Multiplicative conditional feedback systems are believed to represent an entirely new class of systems.

A HEURISTIC DESCRIPTION OF THE MULTIPLICATIVE SYSTEM

A simplified multiplicative conditional system is shown in Fig. 7. Three computing devices have been introduced in this configuration: a divider, a power-raising device, and a multiplier. Consider the system as a regulator when r is a positive constant and the low frequency gain of G is unity. In the absence of a disturbance, the following signal relations exist

$$c=r, \frac{r}{c}=1, \frac{r^{n+1}}{c^n}=r \qquad (20)$$

If the low-frequency gain of G changes from unity to k, where k represents a new steady-state gain such as would result from a change in load, a new system steady state will arise in which the new signal relations are

$$c=r^{n+1}\sqrt{k}, \text{ and } \frac{r}{c}=\frac{1}{n+1\sqrt{k}} \qquad (21)$$

Hence, multiplicative feedback has effectively altered the gain to a value $n+1\sqrt{k}$. This behavior is analogous to the reduction of an additive disturbance in the conventional feedback loop by a factor $1/(1+k)$, where k is the low-frequency loop gain.

When the system is used as a servo, it is necessary to introduce a reference model B. Such a system is shown in Fig. 8, where $B(s)=G(s)$ as in the additive

system. Further, let the low frequency gain of B and of G be unity. In the absence of a disturbance, the following relations hold provided r and c are positive

$$\frac{v}{c}=1, r\frac{v^n}{c^n}=r, C=GR, \text{ and } V=BR \qquad (22)$$

Since the over-all system is linear in the absence of disturbances, the Laplace transform may be used to describe the input-ouput relationship.

Assume now that a disturbance occurs changing the low frequency gain of G to k. In the ensuing steady state, v/c will have a constant value because signal variations in v are duplicated by signal variations in c. Since v/c is a constant, the system again has a linear input-output response relationship, and

$$C=R^{n+1}\sqrt{k}G \text{ and } \frac{v}{c}=\frac{1}{n+1\sqrt{k}} \qquad (23)$$

The question now arises as to the behavior of the system during the transient interval. At the present time, it may be said that, for all signals and disturbances having frequency components that are confined to a frequency range over which G is essentially a constant, the resultant signal c can be represented as the product of two functions: the output c caused by the disturbance if r is maintained at unity, and the output c caused by the signal alone. All questions pertaining to the stability of the multiplicative loop cannot be answered at this time. The following section contains an example of typical modes of behavior for simple multiplicative feedback systems.

RESPONSE OF A SIMPLE MULTIPLICATIVE SERVO

Consider the servo system shown in Fig. 7. In the light of the equivalence shown in Fig. 3, the configuration in Fig. 7 is the equivalent of a classical system employing a ratio comparator. This system has interesting modes of behavior. Suppose G has the form

$$G=\frac{K}{1+Ts}; \quad T=24 \text{ seconds, } K=1 \qquad (24)$$

Let r and c initially have the steady value $r=c=1$. The incremental response in c to an additive step in r is shown in Fig. 9 for various values of the exponent n.[6,7] For comparison, the step response of G is shown. From the system of Fig. 7 it is readily shown that the differential equation defining the response c is

$$\frac{T}{n+1}\frac{d}{dt}(c^{n+1})+c^{n+1}=r^{n+1} \text{ for } c>0 \qquad (25)$$

An equation of the same form governs the response of the system to a multiplicative disturbance such as a sudden decrease in the gain factor K of G. Equation 25 shows that response in the variable c^{n+1} has the time constant

$$T_{n+1}=\frac{T}{n+1} \qquad (26)$$

The exponent n in a multiplicative servo is equivalent to the loop gain of a classical feedback system employing a subtractive comparison. This result as well as that given in equation 21 provides a simple example of the use of multiplicative feedback to alter and calibrate the transmission characteristics of a transducer.

The nonlinear behavior illustrated in the foregoing is produced intentionally by employing essentially nonlinear system components. However, almost all practical control systems contain nonlinearities that are natural to the components and generally undesired. The concept of conditional feedback alters the significance of many of these residual nonlinearities.

NATURAL NONLINEARITIES IN ADDITIVE CONDITIONAL FEEDBACK SYSTEMS

Such nonlinearities as torque saturation in motors, backlash in gears, stiction on shafts, clipping in amplifiers, and nonlinear gain in synchros cause the designer great trouble. Most of these nonlinearities contribute to the instability of a feedback loop and especially to the instability of the classical feedback system. The reason for this condition may

Lang, Ham—Conditional Feedback Systems

be expressed in an interesting manner with the help of Fig. 3. Fig. 3 shows that the classical feedback system is equivalent to a conditional configuration in which the reference model is ideal. Hence the signal ε in the feedback loop contains components caused by all of the nonlinearities in the loop, the influence of each of which feedback is endeavoring to suppress. The classical feedback system does not permit any form of nonlinearity or any form of imperfect linearity to be fully tolerated; it strives for the ideal and commonly becomes unstable in so doing.

However, it is clear that in many practical applications certain residual nonlinearities in the relation of output to the input are not undesirable. It is an unfortunate limitation of the classical feedback system that such admissible nonlinearities contribute to the instability of the feedback loop. The conditional feedback system, on the other hand, provides a direct means for accepting tolerable nonlinearities in the input-output response and at the same time largely prevents these nonlinearities from influencing the stability of the feedback loop. This remarkable behavior is realized by building into the reference model, B in Fig. 2, the tolerable nonlinearities in the components A, G_1, and H. The desired response v at the output of B then contains components caused by these nonlinearities. These components of v cancel the components of the fed-back signal b caused by the nonlinearities in A, G_1, and H. Hence the stability of the feedback loop, in so far as it is excited by the input signal alone, is unaffected. The conditional configuration does not remove the influence of any nonlinearities on the stability of the feedback loop when excited by an additive output disturbance. Hence some care must be exercised in exploiting the aforementioned input-output characteristics.

An excellent example of the use of a conditional feedback system to overcome a serious instability in a classical system is the following. It is well known that a classical system that is conditionally stable for small input signals may be seriously unstable for large signals which cause an element such as a motor to torque saturate. The instability is caused by a reduction in effective loop gain. If the torque-saturating characteristic is included in the reference model of a conditional system, this instability cannot be excited by input signals.

The basic significance of the foregoing is that, for purposes of input-output response, the conditional feedback system

permits the regulating action of the feedback loop to act in a manner best suited to the particular application. Perhaps no examples of this fact are more illuminating than those that spring from the problem of the human operator in a feedback system.

Application of Conditional Configurations to the Human Operator Problem

Whether it be in an automobile, an aeroplane, a steel mill, or an economic system, the importance of system response characteristics in the presence of human operators is profound.[8] It is recognized that the human operator is described by neither a well-defined nor a linear transmission function. Certain basic characteristics can, however, be distinguished. There is what approximates to a pure time delay in motor response to a sudden, say, visual stimulus. There is a blocking action against stimuli received in too rapid succession. The pattern of response changes as experience in a given environment is accumulated. Conditional feedback systems provide the means for exploiting the basic characteristics of the particular human operator by achieving optimum input-output response with adequate loop stability. The design example given is suggestive of the improvement that can be realized in the presence of pure time delays. By providing a linear or nonlinear representation of the human operator in the reference model, the operator is called upon to behave as this model rather than to behave ideally as in the classical system. If the reference model B in Fig. 2 is considered to have a variable structure, it is apparent that the conditional feedback configuration provides an interesting technique for the determination of approximate describing functions for the human operator. The determination is carried out by adjusting the structure of B until the loop signal ε falls to some rms value.

Conclusions

This paper has introduced the concept and elaborated the basic theoretical and practical implications of conditional feedback. Systems are said to be conditional feedback systems when feedback in these systems is used solely in a regulating role to reduce the influence of disturbances of all types. It is distinctive of linear additive conditional feedback systems that the forms of the input-output responses and of the disturbance-output responses are independent. This fundamental fact permits the realization of a

broad new range of response characteristics, especially in the presence of such classically difficult factors as pure time delay. The design procedure for such systems is unusually direct and simple.

A classification of system disturbances as being basically additive or multiplicative in character has led to the concept and practical implementation of multiplicative feedback systems employing ratio comparisons and multiplicative corrections. A new range of response characteristics is available from these systems.

Conditional feedback systems provide means for accepting certain component nonlinearities in the relation of system output to system input and, at the same time, for removing the influence of these nonlinearities on the stability of the feedback loop as excited by the input signal. This fact alters the significance for the designer of many residual nonlinearities and permits certain modes of instability to be suppressed. Conditional feedback systems, except through the new class of multiplicative systems, make no new contribution to the regulator problem but they provide a basis for improvement for all classes of servo problems and not least for the class involving human operators.

References

1. Theory of Servomechanisms, H. L. Hazen. *Journal*, Franklin Institute, Philadelphia, Pa., vol. 218, Sept. 1934, pp. 279–330.

2. Principles of Servomechanisms (book), G. S. Brown, D. P. Campbell. John Wiley and Sons, Inc., New York, N. Y., 1948.

3. Network Analysis and Feedback Amplifier Design (book), H. W. Bode. D. Van Nostrand Company, Inc., New York, N. Y., 1945.

4. Sampled-Data Control Systems Studied Through Comparison of Sampling with Amplitude Modulation, William K. Linvill. *AIEE Transactions*, vol. 70, pt. II, 1951, pp. 1779–88.

5. Synthesis of Transfer Functions by Active RC Networks with a Feedback Loop, D. B. Armstrong, F. M. Reza. *Transactions*, Institute of Radio Engineers, New York, N. Y. vol. CT-1, June 1954.

6. An Analytical Study of a Multiplicative Feedback System, J. E. S. Stevens. *Thesis*, University of Toronto, Toronto, Ont., Canada, 1954.

7. A Practical Study of a Multiplicative Feedback System, R. J. Kavanagh. *Ibid.*

8. The Human Operator of Control Mechanisms, W. E. Hick, J. A. V. Bates. *Monograph No. 17.204*, United Kingdom Ministry of Supply, London, England, May, 1950.

———————◆———————

Discussion

George C. Newton, Jr. (Massachusetts Institute of Technology, Cambridge, Mass.): The authors are to be commended for this interesting approach to the problem of suppressing disturbances without imposing
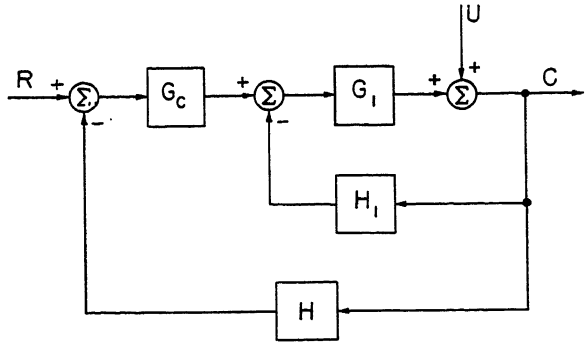
Fig. 10 (left). Parallel feedback configuration

Fig. 11 (right). Single-loop equivalent to parallel feedback configuration
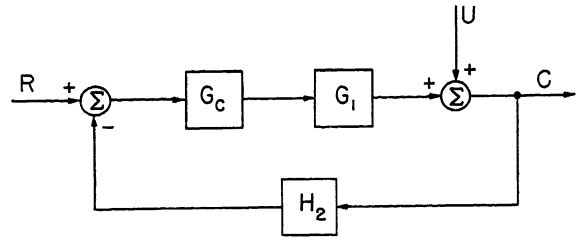
constraints on the input-output transfer characteristic. This is an important problem and can be approached in a number of different ways. The authors' approach is not an unreasonable one in those instances where the input signal is available. Frequently, however, only the difference between the input signal and the primary feedback signal is available with any degree of precision; e.g., a radar tracking system and a position follow-up using synchros for error-sensing. It may be concluded from this paper that the classical single-loop feedback system is unable to do such as can be done with a conditional feedback system configuration. Thus, one may feel at a loss when meeting a problem in which the input signal is not available for filtering prior to summing with the primary feedback signal. It is one purpose of this discussion to show that conventional configurations may be used to achieve results identical with those achieved by the authors. In these configurations there is no need for operating on the input signal.

Consider first the configuration shown in Fig. 10. In this figure, an auxiliary feedback path is used to suppress the effects of disturbances. After the auxiliary feedback transfer function H_1 has been adjusted, the other adjustable transfer functions (usually G_c only) can be adjusted to achieve the desired input-output transmission characteristic. To observe this consider the equation for the output signal

$$C = \frac{G_c G_1 R + U}{1 + G_1 H_1 + G_1 G_c H} \quad (27)$$

This equation compares with equation 5. The behavior of this parallel feedback configuration is identical with the authors' system providing

$$AG_1 = \frac{G_c G_1}{1 + G_1 H_1 + G_1 G_c H} \quad (28)$$

$$1 + G_1 G_2 H = 1 + G_1 H_1 + G_1 G_c H \quad (29)$$

These equations may be solved for G_c and H_1, assuming that A, G_1, G_2, and H are known. The solutions are

$$G_c = A(1 + G_1 G_2 H) \quad (30)$$

$$H_1 = [G_2 - A(1 + G_1 G_2 H)]H \quad (31)$$

Thus, in principle, the authors' system can be realized without need for operations on the input signal. In actual practice, the auxiliary feedback in the case of a positional servomechanism might come from a tachometer attached to the output shaft and H_1 would be adjusted to limit the effect of disturbances. There is no need to use equation 31 since G_2 is arbitrary within wide limits. G_c is then adjusted

to achieve the desired input-output transfer function.

It has been shown that a parallel feedback configuration can be used to suppress disturbances without constraining the input-output transfer functions. Theoretically, it is perfectly possible to realize identical system behavior using a single-loop equivalent to the parallel feedback configuration. Fig. 11 shows such a single loop. If, in this figure, the transmission of the feedback elements is set in accordance with the equation

$$H_2 = H + \frac{H_1}{G_c} \quad (32)$$

the behavior of the single loop will be identical with the behavior of the parallel feedback configuration. Thus, it should not be concluded that single-loop feedback systems are unable to achieve independent adjustment of the disturbance-suppression characteristic and the input-output transmission characteristic. However, in practice, realization of the required transfer functions is usually simpler if a parallel feedback configuration or the configuration suggested by the authors is used.

The authors' application of their conditional principle to multiplicative systems appears to be novel. In connection with process control problems there may be considerable merit in the use of conditional multiplicative feedback configurations. However, a number of questions are left unanswered. For example, in the case of complex G functions, is not the stability sensitive to the output signal level since the loop gain for incremental signals depends on this quantity? Also, are the advantages of the multiplicative system sufficient to justify the use of multiplier, divider, and power-raising devices which are not as easily realized as summing and operational amplifying devices? I look forward to the future publications on the multiplicative systems promised by the authors.

H. C. Ratz (Ferranti Electric Limited, Toronto, Ont., Canada): Consideration of the conditional feedback approach to control problems described by the authors leads naturally to discussion of its relationship to conventional feedback systems. The relationship of conditional systems to classical systems is topologically equivalent to the relationship of bridge circuits to series-parallel circuits. In the conditional system, when the model is chosen for balance, as given by equation 4, there is no signal in the feedback path through G_2 and, hence, the choice of signal input-

output response is independent of the stability of the feedback loop. In Fig. 4, the authors have drawn a classical system with a prefilter which has the same transfer function as the conditional system of Fig. 2. However, in Fig. 4, there is a signal in the feedback path through G_2, the effect of which must be exactly cancelled by the prefilter.

It is possible to draw a conventional feedback system without a prefilter which is linearly equivalent to a conditional system. Fig. 12 shows such a system which is topologically equivalent to Fig. 2. The Laplace transform of the output in Fig. 12 is given by equation 3 with respect to both signal and disturbance. Furthermore, if the model is defined by equation 4, then equations 5, 6, and 7 apply in this case also. Thus, the input-output response is unaffected by any adjustment of G_2 which may be made to alter the character of the disturbance-output response.

There are, however, at least two fundamental differences between this conventional configuration and a conditional system. In Fig. 12, there is always signal in the feedback path so that system stability depends upon the linearity of the components; and the realizability of the transfer functions in the feedback loop is subject to more severe restrictions than in the conditional case. Thus, in special cases, linear equivalents can be formulated, but the contribution of conditional servos is in the simple treatment of nonlinear systems and in the new wide range of feedback configurations which are realizable. Models are already in use for process control problems but the new approach of the authors is to use the model in such a way as to eliminate feedback signals from the error-detecting element. In the conditional system, the feedback loop is employed only to reduce the effect of deviations from model behavior and other disturbances.

H. Tyler Marcy (International Business Machines Corporation, Endicott, N. Y.): It is very appropriate that the authors
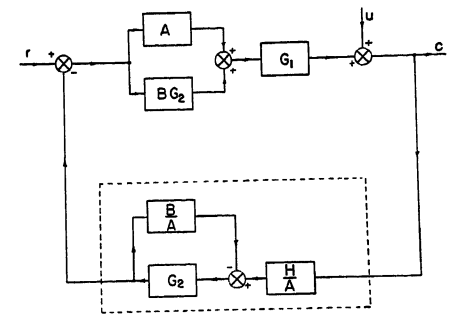
Fig. 12. The conditional feedback system drawn in the configuration of a conventional system

should present this interesting paper at this time. The concept of a control system making use of information contained in the input signal to decrease dynamic errors caused by changes in the signal may be a key technique in many of the more comprehensive process control problems which are receiving much attention today.

There is pertinent literature applying to this subject which makes the word "new" in the paper title subject to question. The concepts of feed-forward have been discussed previously.[1,2,3] It has been widely known that, should it be possible to operate on the input signal, dynamic errors in the control system can be reduced. An important consideration is that feed-forward techniques of error reduction do not change the degree of stability or the natural frequencies of the control system, i.e., the roots of a characteristic equation remain the same. Harris[1] also considered feed-forward as a means for compensating nonlinear factors such as dry friction.

Perhaps the dominant limitation with the use of feed-forward is the nature of the input signal. It must first be possible to measure it directly and then to operate on the signal in a predictive sense. In the presence of noise, this can become impractical. We have, however, learned much about statistical treatment of this sort of problem, since feed-forward was widely discussed some years ago. Regardless of where the disturbance to a control system is applied, predictive knowledge of that disturbance can be employed to decrease dynamic control errors resulting from the disturbance.

REFERENCES

1. THE ANALYSIS AND DESIGN OF SERVOMECHANISM, H. Harris. *Defense Research Council*, Washington, D. C., Sect. D-2, 1942.

2. LINEAR SERVO THEORY, R. E. Graham. *Bell System Technical Journal*, New York, N. Y., vol. 25, Oct. 1946, pp. 616–51.

3. PARALLEL CIRCUITS IN SERVOMECHANISMS, H. Tyler Marcy. *AIEE Transactions (Electrical Engineering)*, vol. 65, Aug.–Sept. 1946, pp. 521–29.

Rufus Oldenburger (Woodward Governor Company, Rockford, Ill.): This valuable contribution to the science of automatic control emphasizes the importance of designing for both servo and regulator performance in many automatic control problems. In the design of speed governors we study the response to speed-setting changes and load rejections, as well as other disturbances. We have found that the design of our governors on the basis of characteristic roots so as to give fast closed-loop response to sudden load changes has also given good response to speed-setting changes.

Unfortunately our most troublesome disturbance is what I like to call the "noise" in the speed signal. This signal is put out by the speed-measuring element. This element corresponds to the authors' feedback element H. As the authors point out, unfortunately this disturbance cannot be reduced without compromising H. The noise can come from gear drive irregularities, prime-mover shaft runout, or other causes. It is a part of the signal one does not wish to respond to, and severely limits the mathematical operations that can be performed on the speed measurement in the computer part of the control.

Prof. James Reswick[1] introduced an auxiliary feedback loop in which the transfer function AG_1 of the forward part is taken to be the transfer function of the extra feedback path, except for a constant of proportionality. If $H=1$ the authors also introduce an element with transfer function AG_1 into an extra loop, but in a different manner by placing it in a forward branch.

The inclusion by the authors of a discussion of nonlinear components is most timely. The American Society of Mechanical Engineers will devote its April 1956 Instruments and Regulators Division conference to nonlinear control. The Russians specialize in this area, as well as in the automatic control field in general; they have a journal devoted entirely to the science of automatic control. The Macmillan Company will soon publish an American Society of Mechanical Engineers book entitled "Frequency Response," which I am editing. This book is to contain leading contributions from all over the world on the various phases of this subject.

REFERENCE

1. DISTURBANCE RESPONSE FEEDBACK—A NEW CONTROL CONCEPT, J. Reswick. *Transactions*, American Society of Mechanical Engineers, New York, N. Y., 1955 (*55-IRD2*).

H. P. Birmingham (Naval Research Laboratories, Washington, D. C.): This important paper is of particular interest to the human engineer concerned with the man as an element in a man-machine control system. A problem under attack is the matter of human performance in pursuit tracking and in compensatory tracking. In compensatory tracking, the human operator is presented only the error (difference between input and output) and manipulates his control on the basis of this information. In the pursuit case, the operator sees the input and his output, or terms proportional thereto, on the same display and is required to keep the difference (error) at a minimum. It has been shown that under some conditions, the human is able to maintain a smaller average error in the pursuit case. It is expected that the authors' analysis can be used to show how the man can turn in this superior performance where the course as well as the error is shown to him, by acting analogously to a conditional system.

G. Lang and J. M. Ham: We greatly appreciate the interest shown by the discussers. Dr. Newton's observation that input signals are not always available with precision is well made but we consider the occasions when input signals cannot be measured with useful accuracy to be rare.

Dr. Newton proceeds to show that in a linear world a conventional feedback configuration, such as shown in Figs. 10 and 11, can be made to have transfer functions identical with those of the basic conditional configuration in Fig. 2. This fact is implied by the prefilter equivalent shown in Fig. 4. As Mr. Ratz has shown in Fig. 12, there are many such linear equivalents. However, we wish to emphasize that our conditional configuration leads to direct realization techniques not readily implemented for general linear equivalents. For example,

Dr. Newton's equations 30 and 31 show how to pick conventional compensating functions G_c and H_1 to achieve equivalence with a conditional configuration. Given the conditional configuration and the freedom to pick A and G_2 independently, equations 30 and 31 are readily used to define G_c and H_1, but we wonder how Dr. Newton would pick G_c and H_1 to give independent input-output and disturbance-output responses without prior reference to a conditional configuration.

In the literature on feedback control systems we are not aware of any collected treatment on the problem of selecting compensating transfer functions to achieve independent conditions on input-output and disturbance-output responses, particularly when nonminimum phase components are present. A practical reason for the difficulty in achieving such conditions is readily discerned from equations 30 and 31. These equations show that any changes in G_c or in H_1 affect both C/R and C/U so that the cut-and-try procedure is commonly resorted to. In this connection, our sample design is exemplifying the directness with which a conditional system design can be realized even under the difficult condition of a pure time delay in the main actuator.

A block diagram for Fig. 10 showing how to realize G_c and H_1 as specified in equations 30 and 31 is given in Fig. 13. If the system is synthesized as shown, it is clear that many more elements are required than for the equivalent conditional system. While it is true that in a linear world G_c and H_1 can be realized with gain and general passive filters, it is not clear how such filters are to be synthesized in practice.

Dr. Newton's remarks about the equivalence of Fig. 11 to the conditional configuration are quite correct but are subject to the qualifications outlined in the foregoing. It should be observed that all of the equivalences discussed by Dr. Newton depend for their validity on linearity. In this connection the remarks of Mr. Ratz are particularly relevant.

With regard to multiplicative systems, Dr. Newton's observation that for complex G functions the loop stability depends on the output signal level is quite correct. To the question of the justification for using multiplier, divider, and power-raising devices in these systems, it may be remarked that multiplicative systems appear to have other than physical worth, namely as model elements for economic systems where per-unit changes are significant. Further, a generalization of the concept of signal comparison leads to insight into the nature of disturbances as suggested in the section entitled "Extension of the Principles to Nonlinear Systems."

Mr. Ratz's comments may be regarded
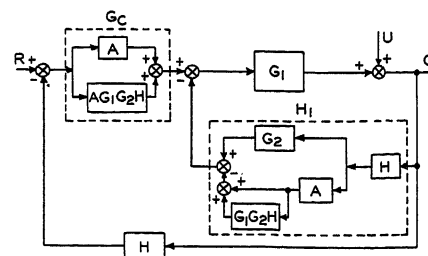
Fig. 13. A synthesis for Fig. 10

as an amplification of the paper and of the remarks of Dr. Newton, and deserve careful reading. Fig. 12 represents the conventional single-loop feedback system equivalent to the conditional configuration. Again it must be noted that, if this system is synthesized as shown in Fig. 12, more elements are required than for the equivalent conditional system.

Mr. Marcy emphasizes the relationship of the conditional feedback systems introduced by us to so-called feed-forward systems to which considerable attention has certainly been given. Perhaps the work of J. R. Moore[1] most closely resembles ours. To the extent that any linear system may be regarded as a combination of feed-forward and feedback elements, the conditional topology introduced by us can be so interpreted. However, the concept and topology of conditional feedback systems employing a reference model has not been described previously in the literature, as far as we know. As Mr. Marcy suggests, the concept of conditional feedback may be particularly illuminating for problems concerning transient process control.

The remarks of R. Oldenburger are interesting. We have heard of Professor Reswick's recent work and are looking forward to seeing his paper.

We appreciate the interest of Dr. Birmingham in the application of conditional feedback systems to the human operator problem. Dr. Birmingham rightly distinguishes the tasks of compensatory and pursuit tracking. These correspond to the regulator and servo definitions as used in the paper. We believe that the conditional topology is significant for the engineering of man-machine systems. A research program on this topic is in progress at the University of Toronto.

REFERENCE

1. COMBINATION OPEN-CYCLE CLOSED-CYCLE SYSTEMS, J. R. Moore. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 39, Nov. 1951, pp. 1421–32.

Design of Control Systems for Minimum Bandwidth

GEORGE C. NEWTON, JR.
MEMBER AIEE

Synopsis: This paper presents a method of designing feedback control systems which minimizes bandwidth for a specified transient error. Reduction of bandwidth is desired in order to attenuate noise, simplify the compensation, and ease the requirements on components operating at high power levels.

Design Specifications

BANDWIDTH TEST

Minimum bandwidth is obtained by adjusting the system weighting function to produce minimum output noise under specially contrived circumstances termed the "bandwidth test." In this paper the bandwidth test is a conceptual scheme for defining bandwidth. During the bandwidth test a stochastic noise signal is used for the control system input. The output of the control system is passed through a filter and its rms value measured. The rms value of the filtered output is correlated with the bandwidth of the control system by comparison with the filtered output of a standard system of any prescribed form but having adjustable bandwidth. The bandwidth of the standard system is adjusted to produce an rms value of its filtered output equal to that of the control system. Both the standard system and the control system are driven from a common noise source during the bandwidth test. Fig. 1 is a block diagram showing this scheme for relating control system bandwidth to its filtered output during a bandwidth test. By definition, the bandwidth of the control system is equal to the bandwidth of the standard system when the rms values of the filtered outputs are equal. In order that minimizing the filtered output of the control system shall be equivalent to minimizing its bandwidth, the standard system together with the noise source and filter must produce a monotonically in-creasing rms output with increasing bandwidth.

PERFORMANCE INDEX

The purpose of any control system is to constrain its output to match a desired output (ideal value) within an acceptable tolerance when acted upon by an input (command) and disturbances. The measure used to assess the agreement between the desired and actual outputs is usually called a performance index. For the design method of this paper the input and desired output are arbitrary transient signals. The integral square of the error between the desired output and the actual output is used as the performance index. In the process of adjusting the control system to minimize its bandwidth, only those weighting functions are used which make the integral-square error equal to or less than a specified value.

SOLUTION FOR WEIGHTING FUNCTION

In the section entitled "A Variational Approach" a general solution is obtained for the weighting function which the control system should have in order to possess minimum bandwidth for a specified integral-square error. This solution is obtained by variational methods. The data necessary to obtain a particular solution are as follows:

In connection with the bandwidth test:

Fig. 1 Scheme for defining of system bandwidth

1. The autocorrelation function (or spectrum) of the noise.

2. The transfer function of the standard system expressed in terms of its bandwidth.

3. The transfer function of the filter.

For the specification of the transient performance:

4. The input.

5. The desired output.

6. The permissible integral-square error.

FILTER

The foregoing discussion explained each of the design specifications except the role of the filter used for the bandwidth test. This filter, used in both the standard and control system channels of Fig. 1, gives control of the rate of cutoff of the transfer function obtained for the control system. Specifically, by making it a differentiator of sufficiently high order the filter "forces" the control system to have a higher cutoff rate than if it were absent.

Examples

To aid in understanding what is implied by the problem statement and the general solution, three examples will be given. All three deal with the problem of designing minimum bandwidth control systems for ramp inputs when the desired output is the same as the input. The first two examples minimize the bandwidth subject to a specification on the integral-square error. It will be shown that the bandwidth required to achieve a given integral-square error increases with the increasing rate of high-frequency cutoff. It will also be shown that the bandwidth is relatively insensitive to variation of the system function from its optimum value.

Because integral-square error as a performance index weights large errors so heavily, it is believed that, under certain conditions, specifying the integral-square error is almost equivalent to specifying the peak error. Suppose the problem of minimizing the integral-square error subject to a specified bandwidth is considered. The solution to this problem is entirely equivalent to the solution for the problem stated in the foregoing. It is suspected that minimizing the integral-square error within a given bandwidth limitation has come close to minimizing the peak error within this same bandwidth limitation.

The third example presents the results of minimizing the bandwidth when the performance index is the peak error. The conditions are the same as for the first example other than for the choice of performance index. It is shown that the peak error obtained is only slightly smaller than in the case of the first example. This third example therefore confirms engineering intuition that specifying the integral-square error is substantially equivalent to specifying the peak error in so far as the bandwidth required for a given peak error is concerned.

Engineering Significance

The design procedure of this paper, which minimizes bandwidth within the constraint of an error specification, differs from conventional methods of synthesis which tend to emphasize error and ignore bandwidth. Why has this new approach been developed? Ideally, good engineering design strives for minimum cost. A minimum cost solution in general does not correspond to either minimum error or minimum bandwidth. Unfortunately, sufficient specifications for a minimum cost design rarely exist. In this situation, conventional methods may result in designs which are unnecessarily expensive because of excessive bandwidth. For example, in focusing attention on errors without regard to bandwidth, there is a tendency to open up the bandwidth so that the transmitted noise is excessive. In the case of mechanical systems, elastances in the gearing and supporting structure may become troublesome if the bandwidth is large. Another result of excessive bandwidth may be saturation of a component which otherwise would be linear. Curing difficulties such as these can be expensive. The tendency is to avoid the difficulties of excessive bandwidth which are latent within conventional techniques, and thereby reduce costs, by using a design procedure which places emphasis on minimizing bandwidth.

Granting the desirability of minimizing bandwidth within the constraint of an error specification, why is integral-square error used for expressing the error tolerance? Integral-square error is adopted because it is analytically convenient and because, under certain conditions as pointed out in the foregoing, it seems to correspond closely to a specification on the peak error. In many problems the peak error is the most important single performance index. The use of integral-square error as a figure of merit for control systems was first made by Hall.[1] Both Hall and more recent investigators[2] have expressed disappointment in the integral-square error as a figure of merit because of the impractical results they have obtained such as a zero damping-ratio solution for a second-order system and excessively oscillatory transients for higher-order systems. One of the interesting features of the approach of this paper is that the impractical results previously revealed in the integral-square error criterion either do not appear or appear to a lesser extent.

In conclusion, it can be stated that

Table I. Summary of Examples

Ex-ample	Given Data					Objective: Find System Function $W_m(\omega)$ Which Minimizes Bandwidth Keeping Equal to or Below a Specified Value the:	Result				
	Noise Spectrum $\Phi_{bvv}(\omega)$	Standard System $W_b(\omega)$	Filter $W_f(\omega)$	Input $v(t)$	Ideal Output $i(t)$		System Function $W_m(\omega)$	Frequency Response of System	Error Response to Ramp Input	Integral-Square Error I_e	Peak Error $ye(t)|_{max}$
1	Constant	$\dfrac{1}{\frac{j\omega}{\omega b}+1}$	1	$\Omega_v t u(t)$	$\Omega_v t u(t)$	Integral-square error	$\dfrac{3\frac{j\omega}{\omega b}+1}{4.5\left(\frac{j\omega}{\omega b}\right)^2+3\frac{j\omega}{\omega b}+1}$	Fig. 3	Fig. 4	$\dfrac{3.375\Omega_v^2}{\omega b^3}$	$0.97\frac{\Omega_v}{b\omega}$
2	Constant	$\dfrac{1}{\left[\left(\frac{j\omega}{\omega b}\right)+1\right]^2}$	$j\omega$	$\Omega_v t u(t)$	$\Omega_v t u(t)$	Integral-square error	$\dfrac{3.63\frac{j\omega}{\omega b}+1}{6\left(\frac{j\omega}{\omega b}\right)^3+6.61\left(\frac{j\omega}{\omega b}\right)^2+3.63\frac{j\omega}{\omega b}+1}$	Fig. 3	Fig. 4	$\dfrac{9\Omega_v^2}{\omega b^3}$	$1.61\frac{\Omega_v}{\omega b}$
3	Constant	$\dfrac{1}{\frac{j\omega}{\omega b}+1}$	1	$\Omega_v t u(t)$	$\Omega_v t u(t)$	Peak error	$\left[\frac{3}{4}\frac{\omega b}{j\omega}-\frac{9}{32}\left(\frac{\omega b}{j\omega}\right)^2\left(1-\epsilon-\frac{8}{3}\frac{j\omega}{\omega b}\right)\right]$	Fig. 6	Fig. 7	Infinit	$0.89\frac{\Omega_v}{\omega b}$

BANDWIDTH TEST SIGNAL

INPUT $b_v(t)$ → CONTROL SYSTEM $w(t)$ → OUTPUT $b_q(t)$ → FILTER $w_f(t)$ → FILTERED OUTPUT $b_f(t)$

NOISE SOURCE

BANDWIDTH TEST CONFIGURATION

INPUT $v(t)$ → CONTROL SYSTEM $w(t)$ → OUTPUT $q(t)$ → Σ → ERROR $y_e(t)$; IDEAL OUTPUT $i(t)$

CONFIGURATION FOR NORMAL OPERATION
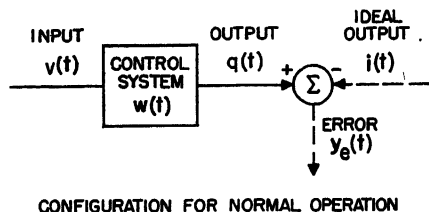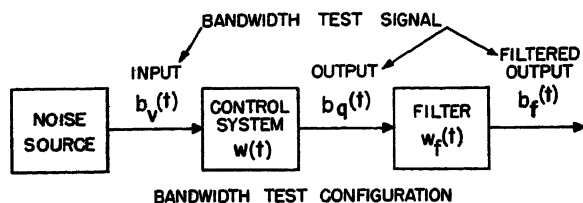
PROBLEM: FIND $w(t)$ WHICH MAKES $\overline{b_f^2(t)}$ A MINIMUM WITHIN CONSTRAINT $\int_{-\infty}^{\infty} dt\, y_e^2(t) \leq I_{em}$

Fig. 2 (left). The variational problem

Fig. 4 (right). System error following ramp input, examples 1 and 2

Fig. 4 graph: NORMALIZED ERROR, $\dfrac{\omega_b}{\Omega_v}\dfrac{y_e}{}$ vs NORMALIZED TIME, $\omega_b t$; EXAMPLE 2, EXAMPLE I

practical considerations may frequently interfere with the realization of control systems which truly minimize bandwidth for a specified performance. It is felt, however, that the procedure discussed in this paper, even if it is not directly applicable, will often serve the designer well by providing bench marks or standards against which he can assess his detailed design arrived at by trial-and-error techniques.

Nomenclature

In general, the nomenclature of this paper is designed to conform where possible with terminology proposed by the AIEE Subcommittee on Terminology and Nomenclature of the Feedback Control Systems Committee in April 1954. In situations where no symbols are provided by the proposed terminology, symbols are chosen which do not conflict. An example of this is the choice of $w(t)$ for the control system weighting function. Usually upper-case letters are used for transforms and lower-case letters for time functions. An exception to this occurs in the case of the function I_{vv}, etc. The use of I_{vv}, etc., as symbols for integrals which are functions of a parameter is dictated by common mathematical usage.

$b_v(t)$ = input during bandwidth test
$b_q(t)$ = output during bandwidth test
$b_f(t)$ = filtered output during bandwidth test
$i(t)$ = ideal transient output (ideal value)
$q(t)$ = output (indirectly controlled variable)
t = time (t_1, t_2, etc., integration variables)
$v(t)$ = transient input (command)
$w(t)$ = control system weighting function
$w_f(t)$ = filter weighting function
$w_b(t)$ = standard system weighting function
$y_e(t)$ = error (system error)
I_e = integral-square error
I_{em} = the specified maximum value of I_e, the integral-square value of the system error $y_e(t) = i(t) - q(t)$

$w_m(t)$ = the system weighting function which minimizes $\overline{b_f^2(t)}$ subject to the constraint that $I_e \leq I_{em}$
ζ = damping ratio of second-order system
τ = shift parameter, integration variable
ω = angular frequency, radians per unit time
$\varphi_{bvv}(\tau)$: see equation 2
$I_{vv}(\tau)$: see equation 3
$I_{vi}(\tau)$: see equation 4
$\Delta(\omega)$: see equation 5
$\Gamma(\omega)$: see equation 5

A Variational Approach

Since, as explained before, minimizing the bandwidth is equivalent to minimizing the noise transmitted by the control system, the system weighting function will be sought which will minimize the rms value of the filtered noise transmitted through the control system within the
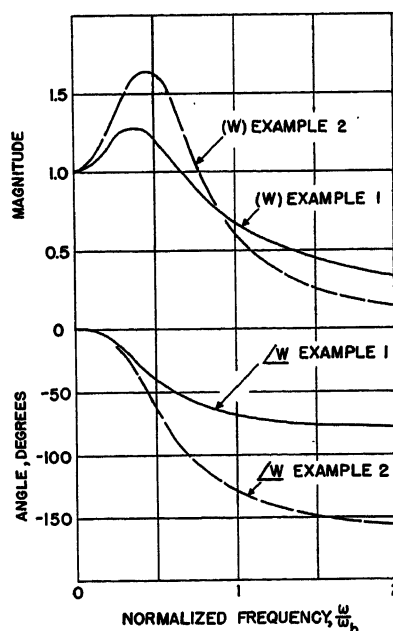
Fig. 3 System functions, examples 1 and 2
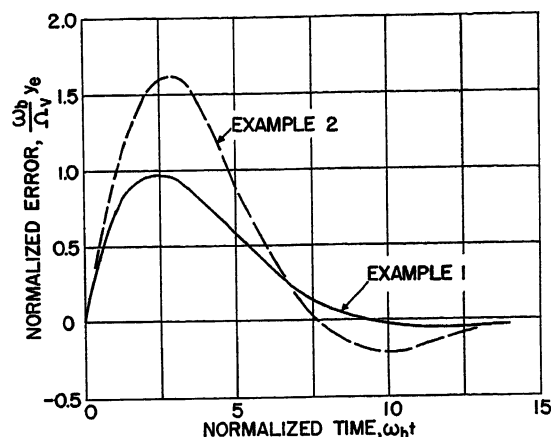
constraint imposed by the specified integral-square error. This is an isoperimetric problem in the calculus of variations.

Fig. 2 is a block diagram illustrating this problem. The details of the solution of the variational problem are presented in the Appendix. The procedure is to formulate the mean-square value of the filtered noise transmitted through the control system and the integral-square error in terms of the system weighting function. By employing the Lagrangian procedure, the isoperimetric problem may be handled as a simple minimization problem. Specifically, the functional, which is the sum of the mean-square value of the filtered noise and a constant times the integral-square error, is minimized by adjustment of the weighting function. The constant is called a Lagrangian multiplier and appears as a parameter in the solution. The solution is obtained in the form of an integral equation by first postulating its existence and then finding the condition it must satisfy for the functional to be stationary with respect to small variations of the weighting function from its optimum value. Either physical considerations or additional analysis may be used to show that this solution produces a minimum and not a maximum.

The problem is now stated more specifically in terms of symbols; see Fig. 2. Given $v(t)$, $i(t)$, I_{em}, $b_v(t)$, and $w_f(t)$, find $w_m(t)$. The solution to the problem is stated in the form of an integral equation

$$\int_{-\infty}^{\infty} dt_4 w_m(t_4) \left[\int_{-\infty}^{\infty} dt_1 w_f(t_1) \int_{-\infty}^{\infty} dt_3 w_f(t_3) \times \varphi_{bvv}(t_1 + t_2 - t_3 - t_4) + \rho I_{vv}(t_2 - t_4) \right] - \rho I(t_{vi2}) = 0 \quad \text{for } t_2 \geq 0 \quad (1)$$

Here

$$\varphi_{bvv}(\tau) \triangleq \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} dt\, b_v(t) b_v(t + \tau) \quad (2)$$

the autocorrelation function of the noise signal used for the bandwidth test. $I_{vv}(\tau)$ is the autotranslation function of the

transient input signal during normal operation. It is analogous to the autocorrelation function for a stochastic signal. It is defined as

$$I_{vv}(\tau) \triangleq \int_{-\infty}^{\infty} dt v(t) v(t+\tau) \qquad (3)$$

Similarly

$$I_{vi}(\tau) \triangleq \int_{-\infty}^{\infty} dt v(t) i(t+\tau) \qquad (4)$$

and is called the cross-translation function of input against the ideal output. ρ in equation 1 is the Lagrangian multiplier which is adjusted to fulfill the condition $I_e \lesssim I_{em}$.

To obtain a solution for $w_m(t)$ from expression 1 requires the solution of an integral equation of the Wiener-Hopf form. In complex problems numerical procedures may be the only feasible approach. If the functions being dealt with are Fourier transformable, Wiener has shown that an explicit solution for $w_m(t)$ may be had in terms of its transform $W_m(\omega)$. This solution is obtained by a process of spectrum factorization. As applied to equation 1 it yields

$$W_m(\omega) = \frac{\left(\dfrac{\Gamma(\omega)}{\Delta^-(\omega)}\right)_+}{\Delta^+(\omega)} \qquad (5)$$

where

$W_m(\omega) \triangleq$ system function, i.e., the Fourier transform of $w_m(t)$

$\Delta(\omega) \triangleq 2\pi \overline{W_f(\omega)} W_f(\omega) \Phi_{bvv}(\omega) + \rho I_{vv}(\omega)$

$\overline{W_f(\omega)}$ denotes the conjugate of $W_f(\omega)$

$\Gamma(\omega) \triangleq \rho I_{vi}(\omega)$

$\Delta^+(\omega) \triangleq$ factor of $\Delta(\omega)$ which has all of the poles and zeros of $\Delta(\omega)$ which lie in the upper half-plane

$\Delta^-(\omega) \triangleq \dfrac{\Delta(\omega)}{\Delta^+(\omega)}$ and has all the poles and zeros of $\Delta(\omega)$ which lie in the lower half-plane

$\left(\dfrac{\Gamma(\omega)}{\Delta^-(\omega)}\right)_+ \triangleq$ component of $\dfrac{\Gamma(\omega)}{\Delta^-(\omega)}$, which has all of its poles in the upper half-plane, such that $\dfrac{\Gamma(\omega)}{\Delta^-(\omega)} - \left(\dfrac{\Gamma(\omega)}{\Delta^-(\omega)}\right)_+$ has all of its poles in the lower half-plane

$W_f(\omega) \triangleq$ system function of filter used in the bandwidth test, i.e., the Fourier transform of $w_f(t)$

$\Phi_{bvv}(\omega) \triangleq$ power density spectrum of the noise used for the bandwidth test. This is $1/(2\pi)$ times the Fourier transform of $\varphi_{bvv}(t)$

$I_{vv}(\omega) \triangleq$ Fourier transform of $I_{vv}(\tau)$

$I_{vi}(\omega) \triangleq$ Fourier transform of $I_{vi}(\tau)$

Here the Fourier transform $F(\omega)$ of a time function $f(t)$ is taken as

$$F(\omega) \triangleq \int_{-\infty}^{\infty} dt \epsilon^{-j\omega t} f(t) \qquad (6)$$

So far the system weighting function $w_m(t)$ or its transform $W_m(\omega)$ has been determined which minimizes the mean-square value (or rms value) of the filtered output during the bandwidth test while

holding the integral-square error for normal operation equal to (or below) a prescribed value. To determine what bandwidth is associated with this minimum mean-square value of the filtered output, the bandwidth of the standard system used for comparison in the bandwidth test is set to a value which produces the same rms filtered output as the control system. Then, by definition, the control system has a bandwidth equal to that of the standard system. This is the minimum bandwidth which the control system may have and still meet the specification on the integral-square error because it has been postulated that the rms filtered output of the standard system is a monotonically increasing function of its bandwidth.

The preceding analysis has considerable flexibility in its definition of bandwidth. Three factors must be specified before a numerical determination of bandwidth is possible: the nature of the noise source, the transmission characteristic of the standard system, and the system function of the filter; see Fig. 1. In solving the examples to be given, a white noise source is always used. The standard system is always taken to be of the form

$$W_b(\omega) = \frac{1}{\left(\dfrac{j\omega}{\omega_b}+1\right)^\alpha} \qquad (7)$$

This binomial form seems to agree better with the intuitive concept of bandwidth than an ideal, sharp cutoff filter. For the filter in the examples to be given a system function corresponding to differentiation is always used, i.e.

$$W_f(\omega) = (j\omega)^{\alpha-1} \qquad (8)$$

The parameter α controls the rate of cutoff of the control system: i.e., the cutoff rate of the control system will be 3α decilogs (6α decibels) per octave. The filter and the parameter α were introduced to give the designer independent control of the cutoff rate. Without these, a minimum bandwidth system always will exhibit the least possible cutoff rate, which is 3 decilogs per octave for the conditions used in the examples.

It should be noted that the specification of the standard system does not affect the solution for the control system since it enters into the problem after the solution has been made, not before. This is in contrast to filter specification which does affect the solution for the control system. The specification of the standard system merely affects the numerical value of the bandwidth assigned to the control system. However, in specifying the standard system care must be taken to make it consistent with the filter employed. If a white noise source is used for the bandwidth test, the product of the filter and standard system functions must cutoff at least as fast as $1/\omega$ to ensure a finite rms value for the filtered output from the standard system channel of Fig. 1.

Before passing to the examples, it should be noted that the solution for the system function given by equations 1 or 5 can be considered equally well as that which minimizes the integral-square error within a specified bandwidth. This viewpoint is reflected in certain graphs for the examples which have been normalized with respect to bandwidth rather than integral-square error. Also, it should be pointed out that equations 1 and 5 represent solutions to the minimum bandwidth problem in its simplest form. More complex formulations are possible. For example, it has been assumed in this paper that the complete control system is under the designer's control. If a portion of the control system is fixed, the problem can be reformulated and a solution can be obtained which includes the fixed elements as part of the given data. Likewise, only transient inputs are discussed in order to reduce the length of this paper. It is equally simple to handle stochastic inputs. Using the methods of reference 4, solutions for these alternate formulations may be devised.

Examples

In this section three examples are presented. The first two illustrate the application of the foregoing general results to a particular problem. The third example is for the purpose of confirming what might be intuitively supposed namely, that minimizing the bandwidth for a specified integral-square error is almost equivalent to minimizing the bandwidth for a specified peak error under the conditions of example 1.

Table I summarizes the examples used to illustrate the design technique discussed in the preceding section. All three examples employ white noise for the noise source used in the determination of bandwidth, and all three consider the normal input to be a ramp function. The ideal output for the three cases is assumed to be identical with the input. In the analysis given in the foregoing the transient functions are defined throughout an infinite region in time. Therefore there is no such thing as an initial condition. Initial conditions in conventional analysis simply account for the effect of inputs prior to $t=0$. The standard system for examples 1 and 3 is a low-pass binomial

filter of the first order. For example 2 a binomial filter of the second order is used as the standard system. The filter employed in the bandwidth termination for examples 1 and 3 is simply a constant; for example 2 it is a simple differentiator. In the case of the first two examples the objective of the design is to find the system function which minimizes the bandwidth subject to a specification on the integral-square error. For example 3, on the other hand, the system function is sought which minimizes the bandwidth within the constraint imposed by specification of the peak error.

Examples 1 and 2 are solved using the explicit solution for the system function given by equation 5. Unfortunately, space does not permit giving the details of these solutions here. The reader may question how solutions can be obtained using equation 5 when the transform for the autotranslation function of the input ramp does not exist. Convergence of the transform can be assured either by employing a convergence factor as part of the transform definition or by considering the solution to be the limit for an exponentially damped ramp as the damping approaches zero. After the solution for the system function has been obtained it is necessary to evaluate the integral-square error and the mean-square value of the noise during the bandwidth test. The table of definite integrals given in the Appendix of reference 5 is often useful here.

Because in example 3 the objective is to minimize the bandwidth subject to a specification of the peak error rather than the integral-square error, the procedures given in the preceding analysis and hence equation 5 are not applicable. In fact,
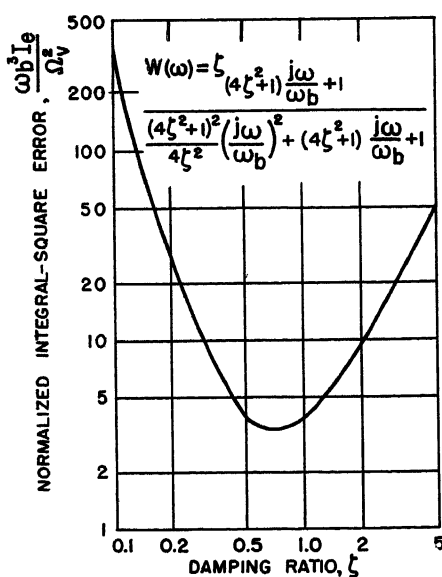
there is no general approach to problems which involve quantities like peak error. Therefore, in handling this example it is necessary to employ approximate methods. By assuming the error is always positive a modification of the Ritz[6] method of solving the variational problem may be used. By taking successive approximations a trend can be established and a guess can be made as to the form of the weighting function which would solve this problem. Next a test was made of this guess to see if it did indeed fulfill the requirements. The net result was that the guess did prove to be the correct one. It further appeared that the initial assumption that the error is always positive was not material, since a change in the system function in any direction which would tend to produce positive, or positive and negative, error causes an increase in the bandwidth of the system for the specified peak error.

The results are now compared for examples 1 and 2. Since the objective is to minimize the bandwidth subject to a specified value of the integral-square error, it is known in advance that the system functions must be such that the steady-state error for a constant velocity is zero. Inspection of the system functions presented in Table I show that they meet this requirement. In the first example, the filter used in the bandwidth test is simply unity. In order to produce a finite bandwidth it is necessary for the system function to cut off at some rate at the higher frequencies. Inspection of the system function for example 1 shows that it cuts off at the rate of 3 decilogs per octave. On the other hand, in example 2 the filter used for bandwidth determination is a simple differentiator. This increased emphasis of the high frequencies results in the system function's cutting off at the rate of 6 decilogs per octave. This illustrates the way in which the designer can control the rate of cutoff. But notice what expense this increased rate of cutoff has imposed. The integral-square error is more than $2\frac{1}{2}$ times as great for a given bandwidth in the case of example 2 than in example 1. Stated another way, in the case of these two examples, almost 40 per cent more bandwidth is required to accommodate a 6-decilog-per-octave cutoff rate than would be necessary to accommodate a 3-decilog-per-octave rate. Inspection of Fig. 3 reveals another price which has been paid to achieve a higher rate of cutoff. Fig. 3 shows the frequency response of the system functions for examples 1 and 2. Notice the higher resonant peak in the magnitude characteristic of exam-

ple 2 relative to example 1. Fig. 4 shows the error as a function of time following the application of the ramp. Initially both the systems for examples 1 and 2 exhibit errors building up at the rate of the ramp, as was expected. The low cutoff rate system of example 1 is able to start reducing the error somewhat sooner than the higher cutoff rate of example 2. Note also that the error response for example 2 shows a greater overshoot than is exhibited by example 1; this ties in with the larger peak magnification of example 2.

It is interesting to compare examples 1 and 2 with Lathrop and Graham's[7] examples Table I, case 2, and Table II, case 2. Using integral-time-absolute error as their performance index, they find that ideally certain parameters should approach infinity. Making these parameters approach infinity effectively makes their systems have infinite bandwidth. If Lathrop and Graham had been able to incorporate a bandwidth specification or had minimized bandwidth holding their performance index constant, they would not have been troubled with parameters tending toward infinity.

How sensitive is bandwidth to variations of a system function from its optimum value? Or alternatively for a given bandwidth, how sensitive is the integral-square error to a departure of the system function from its optimum form? Fig. 5 is designed to answer this question. Since the system function of example 1 is second order, it may be characterized in terms of a damping ratio. The optimum system function has a damping ratio of $\sqrt{2}/2$. Fig. 5 is a plot of the normalized integral-square error as a function of damping ratio ζ as it is varied over a considerable range. Notice that the integral-square error increases rather slowly at first for damping ratios in the range 0.5 to 1.0, but quite rapidly when the damping ratio is much outside this range.

Fig. 5. **Effect of damping ratio variation, example 1**
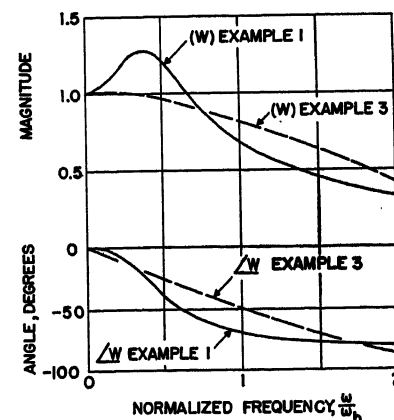
Fig. 6. **System functions, examples 1 and 3**
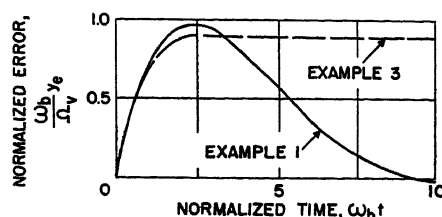
Fig. 7. System error following ramp input, examples 1 and 3

Examples 1 and 3 are now compared. The only difference between these two examples is the performance index which is specified while the bandwidth is minimized. The specification in example 1 was on the integral-square error whereas in the case of example 3 it was on the peak error. In both examples the system functions exhibit a 3-decilog-per-octave cutoff rate. In example 3 there is a steady-state velocity error, which is quite permissible in view of the fact that a specification of the peak error in no way bounds the steady-state error except that it shall be equal to or less than the peak error. This is in distinct contrast to example 1, where the specification on the integral-square error definitely constrained the steady-state error to be zero. Fig. 6 shows the frequency response of the systems for examples 1 and 3. Note that the phase characteristic for example 1 starts off with zero slope corresponding to the zero steady-state velocity error, whereas in example 3 the phase characteristic starts off with a finite slope. Fig. 7 shows the error response following the application of the ramp input for these two examples. Observe the similarity between the initial portions of the error response of example 1 and that of example 3. Notice further that for a specified bandwidth the peak error of example 3 is only slightly smaller than in the case of example 1. What is most spectacular about these two examples is the fact that a 9-per-cent increase in the peak error from that of example 3 gives the zero velocity-error system of example 1, the bandwidth being held constant. On the basis of this result it is concluded that a specification on the integral-square error is almost equivalent to a specification on the peak error under the conditions of these examples, i.e., almost equivalent from the point of view of minimizing bandwidth for a given peak error.

Conclusions

This paper has demonstrated the feasibility of designing control systems to minimize bandwidth. A general solution to the problem of minimizing bandwidth under the condition of a specified integral-square error for a transient input has been presented. The nature of the results to be expected from this general solution has been indicated by means of examples.

The solution for the control system weighting function (or system function) determined by the variational procedure of this paper accomplishes only part of the total design task. No discussion of the realization of a system function by means of closed-loop configurations using available components has been given; this matter alone could well form the subject of a series of technical papers. What has been done is to indicate one approach to establishing a system function which can act as a design goal for conventional system design techniques.

The solution of this paper represents but one application of variational methods[8] to electrical engineering design problems. The author believes that variational techniques have much to offer our branch of engineering and therefore should receive greater emphasis from both teachers and practicing engineers.

Appendix. Derivation of Basic Integral Equation

First, find the mean-square value of the filtered noise during the bandwidth test as a function of $w(t)$. Referring to Fig. 2, the convolution integral is used to write

$$b_f(t) = \int_{-\infty}^{\infty} dt_1 w_f(t_1) b_q(t-t_1) \qquad (9)$$

$$b_u(t-t_1) = \int_{-\infty}^{\infty} dt_2 w(t_2) b_v(t-t_1-t_2) \qquad (10)$$

Therefore

$$b_f(t) = \int_{-\infty}^{\infty} dt_1 w_f(t_1) \int_{-\infty}^{\infty} dt_2 w(t_2) b_v(t-t_1-t_2) \qquad (11)$$

The square of $b_f(t)$ can be written as

$$b_f^2(t) = \int_{-\infty}^{\infty} dt_1 w_f(t_1) \int_{-\infty}^{\infty} dt_2 w(t_2) \int_{-\infty}^{\infty} \times$$
$$dt_3 w_f(t_3) \int_{-\infty}^{\infty} dt_4 w(t_4) b_v(t-t_1-t_2) b_v(t-t_3-t_4) \qquad (12)$$

In equations 9 through 12 t_1, t_2, t_3, and t_4 are integration variables. Since $b_v(t)$ is a stochastic function of finite average power, it possesses an autocorrelation function defined as

$$\varphi_{bvv}(\tau) \triangleq \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} dt\, b_v(t) b_v(t+\tau) \qquad (13)$$

The mean-square value of $b_f(t)$ is by definition

$$\overline{b_f^2(t)} \triangleq \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} dt\, b_f^2(t) \qquad (14)$$

Integrating and averaging both sides of equation 12 (with an interchange of order on the right side) yields

$$\overline{b_f^2(t)} = \int_{-\infty}^{\infty} dt_1 w_f(t_1) \int_{-\infty}^{\infty} dt_2 w(t_2) \int_{-\infty}^{\infty} \times$$
$$dt_3 w_f(t_3) \int_{-\infty}^{\infty} dt_4 w(t_4) \varphi_{bvv}(t_1+t_2-t_3-t_4) \qquad (15)$$

This equation gives the mean-square value of b_f as a functional of $w(t)$.

Next, the integral-square error during normal operation as a function of $w(t)$ is sought. From Fig. 2

$$y_e(t) = i(t) - q(t) \qquad (16)$$

$$y_e^2(t) = i^2(t) - 2i(t)q(t) + q^2(t) \qquad (17)$$

Again employing the convolution integral

$$q(t) = \int_{-\infty}^{\infty} dt_2 w(t_2) v(t-t_2) \qquad (18)$$

$$q^2(t) = \int_{-\infty}^{\infty} dt_2 w(t_2) \int_{-\infty}^{\infty} dt_4 w(t_4) v(t-t_2) \times$$
$$v(t-t_4) \qquad (19)$$

Recall the definition of the autotranslation functions

$$I_{ii}(\tau) \triangleq \int_{-\infty}^{\infty} dt\, i(t) i(t+\tau) \qquad (20)$$

$$I_{vi}(\tau) \triangleq \int_{-\infty}^{\infty} dt\, v(t) i(t+\tau) \qquad (21)$$

$$I_{vv}(\tau) \triangleq \int_{-\infty}^{\infty} dt\, v(t) v(t+\tau) \qquad (22)$$

Integrating both sides of equation 17 after substituting expressions 18 and 19 for $q(t)$ and $q^2(t)$ respectively yields as an expression for the integral-square error

$$I_e = I_{ii}(0) - 2 \int_{-\infty}^{\infty} dt_2 w(t_2) I_{vi}(t_2) + \int_{-\infty}^{\infty} \times$$
$$dt_2 w(t_2) \int_{-\infty}^{\infty} dt_4 w(t_4) I_{vv}(t_2-t_4) \qquad (23)$$

An interchange in the order of integration and use of expressions 21 and 22 were made in deriving this equation, which gives I_e as a function of $w(t)$.

Using the method of Lagrange, the functional is minimized

$$F = \overline{b_f^2(t)} + \rho I_e \qquad (24)$$

where ρ is the Lagrangian multiplier, in order to minimize $\overline{b_f^2(t)}$ subject to the constraint that $I_e \leq I_{em}$.

To do this let

$$w(t) = w_m(t) + \epsilon w_e(t) \qquad (25)$$

where $w_m(t)$ is the solution and $w_e(t)$ is any arbitrary weighting function. Both $w_m(t)$ and $w_e(t)$ must be realizable in the sense that they are zero before $t=0$. Then the condition for $w_m(t)$ to be the solution develops from setting the first derivative of F with respect to ϵ equal to zero at $\epsilon=0$. By equation 24

$$\left.\frac{dF}{d\epsilon}\right|_{\epsilon=0} = \left.\frac{d[\overline{b_f^2(t)}]}{d\epsilon}\right|_{\epsilon=0} + \rho \left.\frac{dI_e}{d\epsilon}\right|_{\epsilon=0} \qquad (26)$$

From equation 12

$$\frac{d[\overline{b_f^2(t)}]}{d\epsilon} = \int_{-\infty}^{\infty} dt_1 w_f(t_1) \int_{-\infty}^{\infty} dt_2 \frac{dw(t_2)}{d\epsilon}$$
$$\int_{-\infty}^{\infty} dt_3 w_f(t_3) \int_{-\infty}^{\infty} dt_4 w(t_4) \varphi_{bvv}(t_1 +$$
$$t_2-t_3-t_4) + \int_{-\infty}^{\infty} dt_1 w_f(t_1)$$
$$\int_{-\infty}^{\infty} dt_2 w(t_2) \int_{-\infty}^{\infty} dt_3 w_f(t_3)$$
$$\int_{-\infty}^{\infty} dt_4 \frac{dw(t_4)}{d\epsilon} \varphi_{bvv}(t_1+t_2-t_3-t_4) \qquad (27)$$

Substituting the right side of equation

25 for $w(t)$ and making use of the fact that $\varphi_{bvv}(\tau)$ is an even function makes it possible to write

$$\frac{d\overline{[b_f{}^2(t)]}}{d\epsilon}\bigg|_{\epsilon=0} =$$

$$2\int_{-\infty}^{\infty}dt_1 w_f(t_1)\int_{-\infty}^{\infty}dt_2 w_e(t_2)\int_{-\infty}^{\infty}dt_3 w_f(t_3)$$

$$\int_{-\infty}^{\infty}dt_4 w_m(t_4)\varphi_{bvv}(t_1+t_2-t_3-t_4) \qquad (28)$$

Similarly

$$\frac{dI_e}{d\epsilon}\bigg|_{\epsilon=0} =$$

$$2\int_{-\infty}^{\infty}dt_2 w_e(t_2)\int_{-\infty}^{\infty}dt_4 w_m(t_4)I_{vv}(t_2-t_4)-2\times$$

$$\int_{-\infty}^{\infty}dt_2 w_e(t_2)I_{vi}(t_2) \qquad (29)$$

Substituting expressions 28 and 29 into equation 26 · for $(dF/d\epsilon)|_{\epsilon=0}$ setting

$(dF/d\epsilon)|_{\epsilon=0}=0$, and rearranging the orders of integration yields

$$2\int_{-\infty}^{\infty}dt_2 w_e(t_2)(\text{left side of equation 1})=0 \qquad (30)$$

Since $w_e(t)$ is arbitrary and not zero for $t \geqslant 0$, the parenthesis must be zero in this range. This completes the derivation of the basic integral equation.

References

1. THE ANALYSIS AND SYNTHESIS OF LINEAR SERVOMECHANISMS (book), A. C. Hall. The Technology Press, Cambridge, Mass., 1943.

2. THE SYNTHESIS OF "OPTIMUM" TRANSIENT RESPONSE: CRITERIA AND STANDARD FORMS, Dunstan Graham, R. C. Lathrop. AIEE Transactions, vol. 72, pt. II, Nov. 1953, pp. 273–88.

3. EXTRAPOLATION, INTERPOLATION, AND SMOOTH, ING OF STATIONARY TIME SERIES (book), N. Wiener. John Wiley and Sons, Inc., New York N. Y., 1949.

4. COMPENSATION OF FEEDBACK CONTROL SYSTEMS SUBJECT TO SATURATION, George C. Newton, Jr. Journal, Franklin Institute, Philadelphia, Pa., vol. 254, Oct. and Nov. 1952, pp. 281–96 391–413.

5. THEORY OF SERVOMECHANISMS (book), H. M. James, N. B. Nichols, R. S. Phillips. McGraw-Hill Company, Inc., New York, N. Y., 1947.

6. METHODS OF APPLIED MATHEMATICS (book), F. B. Hildebrand. Prentice-Hall, Inc., New York N. Y., 1952.

7. THE TRANSIENT PERFORMANCE OF SERVO-MECHANISMS WITH DERIVATIVE AND INTEGRAL CONTROL, Richard C. Lathrop, Dunstan Graham. AIEE Transactions, vol. 73, pt. II, Mar. 1954, pp. 10–17.

8. THE INTRODUCTION OF CONSTRAINTS INTO FEEDBACK SYSTEM DESIGNS, John H. Westcott. Transactions of the IRE, CT-1, Institute of Radio Engineers, New York, N. Y., Sept. 1954, pp. 39–49.

Discussion

A. J. Bargeski (Sperry Gyroscope Company, Great Neck, N. Y.): Mr. Newton's paper is an interesting and much-needed advance in the application of statistical techniques to control system design. Although the use of constraints on the system function has been introduced in other papers, this paper to my best knowledge is the first to realize that some sacrifice in system performance may be necessary if derived bandwidth requirements are to be compatible with those obtainable from functional components available to the system designer. The specification of an upper bound on the integral-squared error rather than its minimization should do much toward reducing the impractical results that are generally derived by the straightforward application of statistical functions.

The limited results the author presents are informative. A comparison of examples 1 and 2 indicates that systems with second-order characteristic equations are the most efficient with respect to ramp function inputs under the given noise conditions. On this basis, it may be possible to construe that single-order systems will suffice for step function inputs, and third-order systems are necessary for constant acceleration inputs. Thus, on an open-loop basis, the order of the zero pole of the response need be no greater than that necessary to provide a zero error for a given input as determined by the classical theory. The addition of the constraint then defines systems in accord with specifications set down by less powerful analytic tools. With the general form of system functions determined by the foregoing technique, the process of minimization with included constraints can be carried out in the manner of reference 5 with respect to the system parameters that can be varied.

The first example specifies an open-loop response having a second-order zero pole. Since lags are inherently present in system functions supplying specified leads, the results of example 2 appear to be the more practicable and thus should be used as the minimization standard. The system function of the second example also agrees with our intuitive concept as to the ideal function since the additional 3-decilog-per-

octave high-frequency fall-off should reduce the noise output. System performance should not be unduly compromised if the lag corner frequency occurs sufficiently far from the zero-decibel point. The inconsistency here undoubtedly lies in the manner of specification of the system bandwidth W_D.

The obvious extensions to this paper have already been mentioned by the author. If these were to be carried out and formalized, the practical design of optimum system functions would be considerably advanced.

Dr. S. S. L. Chang (New York University, New York, N. Y.): The author is to be congratulated for his very sophisticated analysis of the problem as well as for the very useful result he obtained. His result has the flexibility to meet practical requirements. For instance, a designer may attach varying degrees of importance to high-frequency noise (or cost of extended frequency range) by varying the filter response function $W_f(\omega)$. The cost could include fix cost as well as maintenance cost, such as the increased wear of moving components which invariably accompanies wide-open bandwidth.

His result is applicable not only to inputs of a known time function but also to cases in which only some statistical property such as the power spectrum of the input is known. This applicability is inherent in his use of integral-square error criterion.

The minimum bandwidth philosophy of servo design has its roots in the literature at least as far back as 1947 when reference 5 was published. In that book the minimizing over-all error in the presence of noise is analytically equivalent to Dr. Newton's minimizing bandwidth. However, analysis in the book is restricted to determining the coefficients of a certain form of system function.

I have been studying the more general problem by means of frequency domain analysis. For a linear servo, my work gives the same result as Dr. Newton's and will be given here as an alternative to Dr. Newton's analysis at the end of this discussion. For a nonlinear servo my work indicates that in the presence of noise the servo response to signal actually slows down

if the bandwidth is unduly increased. This fact further corroborates Dr. Newton's antihigh-fidelity philosophy of servo design.

For the linear case, it is well known that the integral square of a time function is the same as the integral square of the magnitude of its Fourier transform. The problem is simply to determine $W(\omega)$ subject to the following conditions

$$\int_0^{\infty} W_f(\omega)W(\omega)\bar{W}_f(\omega)\bar{W}(\omega)\Phi_{bvv}(\omega)d\omega$$
$$=\text{minimum} \quad (31)$$

$$\frac{1}{2\pi}\int_0^{\infty}[V(\omega)W(\omega)-i(\omega)][\bar{V}(\omega)\bar{W}(\omega)-$$
$$\overline{i(\omega)}]d\omega=\text{constant} \quad (32)$$

In equation 32, $v(\omega)$ and $i(\omega)$ are Fourier transforms of $v(t)$ and $i(t)$ respectively. If only the power spectrum of $v(t)$ is known, $v(\omega)$ will have arbitrary or undetermined phase. However, the phase of $i(\omega)$ is correlated to that of $v(\omega)$. Multiplying out equation 32, there results

$$\frac{1}{2\pi}\int_0^{\infty}[I_{vv}(\omega)W(\omega)\bar{W}(\omega)-I_{vi}(\omega)\bar{W}(\omega)-$$
$$\bar{I}_{vi}(\omega)W(\omega)+I_{ii}(\omega)]d\omega=\text{constant} \quad (33)$$

The minimization is subject to the restriction that $W(\omega)$ is physically realizable and stable. Let $W_e(\omega)$ be any physically realizable and stable system function, and e be an infinitesmal quantity. Then for any arbitrary

$$W(\omega)=W_m(\omega)+eW_e(\omega) \quad (34)$$

satisfying equation 32, the integral of equation 31 must be equal to a zero-order term plus a vanishing first-order term plus a positive definite second-order term. Using Lagrange's multiplier, we obtain

$$\int_0^{\infty}W_f(\omega)\bar{W}_f(\omega)\Phi_{bvv}(\omega)[W_m(\omega)\bar{W}_e(\omega)+$$
$$\bar{W}_m(\omega)W_e(\omega)]d\omega+\frac{\rho}{2\pi}\int_0^{\infty}\Big\{[I_{vv}(\omega)\times$$
$$W_m(\omega)-I_{bi}(\omega)]\bar{W}_e(\omega)+[I_{vv}(\omega)\times$$
$$\bar{W}_m(\omega)-\bar{I}_{vi}(\omega)]W_e(\omega)\Big\}dw=0 \quad (35)$$

The condition of physical realizability gives

$$\bar{W}_m(\omega) = W_m(-\omega)$$

$$\bar{W}_e(\omega) = W_e(-\omega)$$

$$\bar{W}_f(\omega) = W_f(-\omega)$$

These relations, together with the well-known properties that $I_{vv}(\omega) = I_{vv}(-\omega)$ $\Phi_{bvv}(\omega) = \Phi_{bvv}(-\omega)$, and $I_{vi}(\omega) = I_{vi}(-\omega)$ convert equation 35 into the following form

$$\int_{-\infty}^{\infty} \left\{ [W_f(\omega)\bar{W}_f(\omega)\Phi_{bvv}(\omega) + \frac{\rho}{2\pi}I_{vv} \times \right.$$

$$\left. (\omega)] W_m(\omega) - \frac{\rho}{2\pi}I_{vi}(\omega) \right\} \bar{W}_e(\omega)d\omega = 0 \qquad (36)$$

Since the system functions must vanish at $\omega = \infty$, the path of integration may be extended from $+\infty$ through a large semicircle on the upper half-plane back to $-\infty$. Thus the path of integration of equation 36 encloses the entire upper half-plane. Since $W_e(\omega)$ has no poles in the lower half-plane, $\bar{W}(\omega) = W(-\omega)$ has no poles in the upper half-plane. A sufficient condition for satisfying equation 36 is simply that the expression in the { } bracket of equation 6 has no poles in the upper half-plane. This is also the necessary condition due to the arbitrariness of $W_e(\omega)$. If the expression in the { } bracket has any poles in the upper half-plane, a $W_e(\omega)$ can be chosen such that the sum of residues do not vanish. Therefore

$$[W_f(\omega)\bar{W}_f\Phi_{bvv} + \frac{\rho}{2\pi}I_{vv}(\omega)] W_m(\omega)$$

$$= \frac{\rho}{2\pi}I_{vi}(\omega) + L(\omega) \qquad (37)$$

where $L(\omega)$ has poles in the lower half-plane only.

Using Dr. Newton's definition of $\Delta^+(\omega)$ and $\Delta^-(\omega)$, and dividing equation 37 by $\Delta^-(\omega)$, there results

$$\Delta^+(\omega)W_m(\omega) = \frac{\rho I_{vi}(\omega)}{\Delta^-(\omega)} + \frac{2\pi L(\omega)}{\Delta^-(\omega)} \qquad (38)$$

In equation 38, note that $\Delta^+(\omega) W(\omega)$ has poles in the upper half-plane only. On the left-hand side of equation 38, the first term has poles in both the upper and lower half-planes and the second term is indeterminate but has poles in the lower half-plane only. Therefore

$$\Delta^+(\omega)W_m(\omega) = \left(\frac{\rho I_{vi}(\omega)}{\Delta^-(\omega)}\right)_+ \qquad (39)$$

Equation 39 is the same as equation 5. Throughout the derivation, I have used Dr. Newton's symbols instead of my own to show the parallelism of the two methods.

Max V. Mathews (Massachusetts Institute of Technology, Cambridge, Mass.): Professor Newton has developed an excellent approach to the design of servomechanisms. One of the most valuable aspects is the presentation of an example of the application of the calculus-of-variations methods to design problems in which these methods can provide the designer with assurance that the best possible design has been obtained within the limits of the fixed conditions imposed by the problem.

Professor Newton approached the servo design problem, from the standpoint that the maximum allowable error is given as a fixed quantity and none of the quantities in the servo are specified. He then proceeded to design a servo which is optimum in the sense that it has a minimum bandwidth. An alternate approach, which is frequently employed either by choice or because of the nature of the problem, is to design a servo which yields minimum error under the constraint of some fixed quantity in the servo. The prime mover is a constraint which frequently is specified. However, if the bandwidth is the specified constraint, then minimization of the error results in the same design as Professor Newton's procedure. The identity of the two design approaches is mentioned in the paper, but is sufficiently important to be emphasized here.

The main limitation to the author's approach is that only linear systems are treated and consideration of nonlinear elements (such as saturating elements), which are inherent in almost all servos, does not seem feasible. This limitation is not unique to Professor Newton's paper, since almost all literature in the servo field deals with linear theory. A design method cannot be entirely satisfactory until nonlinear effects are taken into account. For example, a minimum bandwidth system might have larger signal amplitudes in some fixed component than some system using the same fixed component but with a wider bandwidth. Conceivably, the latter system might be cheaper to build.

George C. Newton, Jr.: I wish to express my appreciation to the discussers for their comments on my paper.

As Mr. Bargeski points out, it is perfectly feasible to design a feedback control system for minimum bandwidth by first selecting a configuration and then adjusting the free parameters to minimize the bandwidth in the manner of reference 5. If a little engineering intuition is employed in the selection of the initial configuration, it is often possible to achieve results by this procedure which are either identical to, or almost as good as, those obtained when the configuration is not initially specified. This comes about because small variations of the impulse response of the control system from its optimum shape do not markedly increase either the bandwidth or the integral-square error since both these quantities are stationary in the vicinity of the optimum shape. In other words, both bandwidth and integral-square error are relatively insensitive to departures from the optimum impulse response providing the departures are not too large.

Because it is not essential for engineering purposes to have exactly the optimum impulse response for a control system, it is frequently possible to arrive at a satisfactory design by a trial-and-error technique using conventional design theory. The virtue of the more elaborate analyses, of which this paper is an example, lies in their providing us with bench marks, or reference points, by which we may gauge the results of the simpler design techniques. Also, these more elaborate methods help us to determine whether or not a set of specifications is internally consistent. Certain specifications may be impossible to realize and yet no amount of trial-and-error analysis can prove their unrealizability. On the other hand, the more elaborate methods along the lines of this paper are able to tell the designer which specifications are realizable and which are not. Knowing when a set of specifications is not realizable can save the designer a great deal of time.

Dr. Chang states that he is working on analyses of nonlinear servomechanisms and that in certain instances he has shown the servo response to signal actually slows down as the bandwidth is increased. It is not difficult to imagine certain situations in which this is true, but undoubtedly Dr. Chang has made a very careful study of this phenomenon. I, for one, am looking forward to the publication by Dr. Chang of this investigation.

In my paper the solution for the minimum bandwidth system function is arrived at through the solution of an integral equation for the impulse response. Dr. Chang points out that it is not necessary to first solve for the impulse response. As he has shown, it is possible to set a problem up in terms of the system functions and solve for the optimum system function without ever leaving the frequency domain. Which approach is better depends upon the purposes the user has in mind. By going to the solution via the time domain we do obtain incidentally an integral equation for the impulse response. In problems where only numerical data are available, it may be desirable to remain in the time domain, and to solve the integral equation by numerical procedures. On the other hand, if one is always sure of dealing with time functions whose transforms are analytic, then the frequency domain solution is obviously the best one to use.

As Dr. Mathews points out, the analysis of my paper is based on the assumption that the control system is linear. The problem of nonlinear system synthesis in any general sense is so much more difficult than the problem treated in my paper that there should be no wonder at the paucity of literature on this subject. Considerable research has been conducted on specialized nonlinear systems and the results of this work are being reported in the literature with increasing frequency.

Dr. Mathews mentions that a minimum bandwidth system may have larger signal amplitudes in certain fixed components than some other system using the same fixed components but having a larger bandwidth. I believe that fixed components for which a minimum bandwidth system yields larger signal levels than some other system are relatively rare. In the most recent example in my experience involving fixed components possessing a nonminimum-phase property, the minimum bandwidth solution yielded far lower signal levels in the fixed elements than would have existed otherwise. In the case of minimum-phase fixed components those employed in feedback control systems usually have a low-pass frequency-response characteristic. In such circumstances minimization of integral-square error, or for that matter other error criteria, without regard to bandwidth and in the absence of noise will result in compensation which is the inverse of the transfer function of the fixed elements (assuming that the desired output is the same as the input). With such compensation the signal levels within the fixed components may become very severe indeed.

Considerations in the Development of a High-Power Rectifier Locomotive

H. S. OGDEN
MEMBER AIEE

General Considerations

HIGH-POWER rectifier locomotive design considerations center, first and foremost, around the traction motor. To compensate for the cost of the rectifier equipment, the d-c traction motor must be of a design that enjoys high production and consequent low cost. To be competitive in road performance with contemporary locomotives powered by modern 25-cycle a-c commutator motors, which have the inherent ability to deliver high power at high speeds, the motor selected for the rectifier locomotive must have comparable characteristics. The peak outputs required per motor for short periods on a passenger locomotive are around 1,500 horsepower. It is the ability to deliver power of this magnitude per axle that permits the modern straight electric locomotive to maintain the high-speed schedules that characterize electrified railroads along the Eastern seaboard. In high-speed passenger service it is desirable that the rectifier locomotive be able to duplicate or exceed the performance of present locomotives equipped with 25-cycle commutator motors.

Extensive stand tests have shown that the modern d-c Diesel-electric locomotive traction motor can deliver these high horsepowers for short periods. The tests also indicate that motor commutation is sufficiently good to permit these high outputs to be attained with the expectation of reasonable commutator and traction motor maintenance.

Wave Shape

Single-phase power, when rectified, does not give direct current in the usually accepted sense of the term. If there is no inductance in the circuit, a pulsating direct current is produced; each pulse being the upper half of a sine wave.

Actually, the lower half of the sine wave has been inverted above the zero axis. If sufficient inductance can be added to the rectifier output circuit, the pulses can be smoothed out and the output made to become a constant d-c voltage. The amount of inductance required, however, is so high as to be impractical. But a reactor of reasonable size can be designed if the ripple in the output current is held to approximately 20 per cent with maximum motor currents.

Any method of smoothing out the d-c wave has the disadvantage of tending to cause severe distortions of the current supplied to the locomotive. This distortion sets up conditions for possible telephone influence which may or may not prove to be serious.

Several curatives are available for improving the primary wave. The usual method used (described in the section "Telephone Interference") consists of connecting a capacitor-resistance filter across the input to the rectifier tubes. A less effective method involves connecting a tuned capacitor-reactor circuit of large capacity across the traction motor and smoothing reactor (SR) combination. Such a circuit is tuned to approximately the second harmonic frequency, but so arranged that the reactor voltage is greater than the capacitor voltage (Fig. 1). With these phase relations, the circuit has the characteristic of improving the traction motor ripple, and at the same time increasing the ripple current taken from the supply transformer. Under one condition of operation at continuous load on the traction motor, the ripple observed in the motor current was 19.6 per cent, whereas the ripple coming out of the supply transformer was 47.6 per cent. At this time a 50-cycle filter current of 135 amperes was required. Although the shape of the primary current wave proved to be a rather good approximation of a sine wave, it still had a sharp corner occurring at the end of the rectifier commutation period, and a rather high value of telephone influence factor (TIF) of 25.8 was obtained. A second harmonic filter current of the amount used here requires a very large kva in capacitor and reactors. For this reason the small

benefit obtained from such a connection proves very costly. A circuit of this nature would give a reduction in the size of the main SR, but analysis indicates that the over-all arrangement would not be economical. An equivalent improvement in motor ripple could be obtained by altering the design of the main SR at a considerable saving in weight and space.

Rectifier Connections

There are two types of rectifier connections which might be used for converting alternating current to direct current. The first is the diametric connection, and the second the bridge or 2-way connection. The principal elements of these two types are shown in Fig. 2.

With the diametric connection, the load circuit is connected between the rectifier anodes and the mid-point of the supply transformer. This connection is economical of rectifier tubes at the expense of transformer weight and cost. A glance at the circuit connections will serve to show that the current flow alternates first in one half of the secondary winding of the transformer and then in the other half, as first one and then the other set of tubes conducts. Since the output voltage of the rectifier is determined by the voltage from the mid-point to one end of the transformer, it is evident that the transformer must be wound for twice the voltage of the output circuit, but each section of the winding carries current for only one-half the time. During the period of rectifier commutation, when the outside terminals of the main transformer secondary are (in effect) short-circuited, the SR must force its current through the load circuit to the center tap of the transformer. At this point the current divides and flows in opposite directions through the two halves of the secondary winding and the rectifier tubes back to the load circuit and the SR. The flow of current during the commutating period of the rectifier is in accordance with the classical theory of current flow in
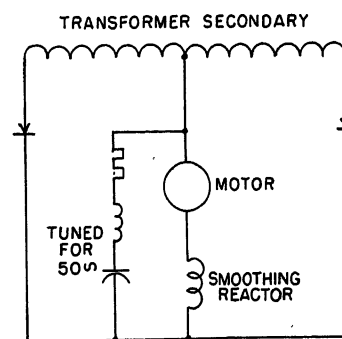
Fig. 1. Second harmonic d-c filter
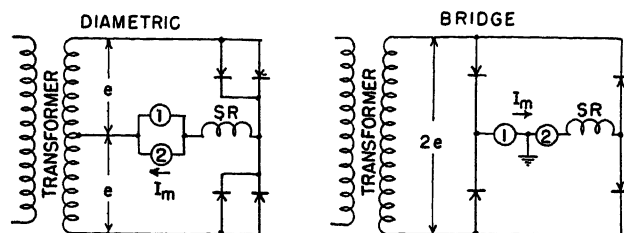
DIAMETRIC   BRIDGE

Fig. 2. Comparison of diametric and bridge rectifier connections

Diametric		Description	Bridge
$\sqrt{2}I_m$	1.	Transformer secondary winding current per motor, continuous	I_m
$2\sqrt{2}I_m e$	2.	Transformer secondary kva	$2eI_m$
e	3.	Volts per rectifier tube	e
$2e$	4.	Inverse volts per tube	$2e$
$2I_m e_t$	5.	Tube loss (e_t =Forward tube drop)	$2I_m e_t$
4	6.	Number of tubes required (same amperes per tube)	4
Requires special bracing for pulsating winding currents	7.	Transformer	standard design
$2I_m(e_m+e_t+e_x)$ (e_x=leakage reactance volts 1/2 transformer secondary)	8.	Main SR driving volt amperes (e_m=motor volts)	$2I_m(e_m+e_t)$
$I_a=\dfrac{1}{1.5\%X}2I_m=1.33\left(\dfrac{1}{\%X}I_m\right)$	9.	Tube arc back current for same per cent reactance transformer (sustained value)	$I_a=\dfrac{1}{\%X}I_m$
$2n$	10.	Number of switches per voltage step	n

that, when the nonconducting tubes start to carry current, the current they carry is a short circuit on the entire secondary of the transformer. The direction of this current is opposite to that of the load current carried by the tube which has been conducting for the previous half-cycle. When the build-up of current reaches a numerical value equal to one-half the value of the current in the previously conducting tube, the two currents flowing in the two halves of the transformer secondary in opposite directions result in zero load current ampere turns, and hence zero primary current. When the short-circuit current becomes exactly equal to the current originally flowing through the previously conducting tube, the algebraic sum of these two currents is zero, and current ceases to flow in the previously conducting tube but does not reverse because of the tube characteristics.

Fig. 2 also shows the elementary connections of the bridge-type rectifier circuit. The tube action and the currents flowing during commutation are much the same as for the diametric connection, except that the SR current does not flow through the main transformer secondary during this period. In addition to forcing the current through the traction motor, the reactor must overcome the drop of two rectifier tubes, whereas in the diametric circuit, only one rectifier tube drop is involved. Because of the leakage reactance of the diametrically connected transformer, and the fact that the instantaneous values of current during the commutating period in the two halves of the main transformer are not always equal, more voltage must be generated by the SR for this connection. The

bridge-connected rectifier therefore requires a smaller physical size of SR than does the diametrically connected rectifier.

In locomotive service, the traction motor voltage must be varied over a rather wide range. This can be done either by taps on the main transformer, or by accelerating resistors connected in series with the motors. In some cases a combination of the two methods may be provided.

Where it is desired to use taps on the main transformer secondary, a diametric connection requires that the voltage on each half of the secondary be balanced at all times. Otherwise, traction motor ripple currents will not be kept within limits.

With the diametric connection, substantial unbalanced stresses are set up in the transformer because of the pulsating nature of the current in the two sections of the secondary as it alternates under the rectifying action of the tubes.

The tabulation included with Fig. 2 shows the relative advantages of the two schemes. This assumes two traction motors which may be connected either in series or parallel. For the diametric connection, the motors are in parallel and the center tap of the transformer is grounded. For the bridge connection, the ground on the traction motors is located between the two motors. This insures that maximum voltages to ground on the traction motors will be the same in both cases. Because of the pulsating nature of the current in the transformer secondary for the diametric connection, the heating effect is equivalent to $\sqrt{2}$ times the current per motor. This means there must be 40 per cent more copper in this secondary winding than in that of the

transformer for the bridge connection. As a result, the transformer as a whole has approximately 20 per cent more kva and 10 per cent more weight. When considering transformers of 5,000-kva capacity, weighing approximately 25,000 pounds, this means a weight saving of 2,500 pounds or better in favor of the bridge circuit. The tabulation also shows that the duty per rectifier tube is the same for the two connections, not considering the fault currents. Item 9 indicates that when a tube arcs back, the arc-back current will be one-third greater for the diametric connection.

The purpose of the SR is to sustain the current flowing through the traction motors during the period of rectifier commutation. For both circuits, the SR has to overcome the voltage drop in the two traction motors and also the forward drop in the rectifier tubes. The diametric circuit requires an SR design of half the voltage and twice the current required by the bridge circuit. In the diametric circuit, however, the SR must force the motor current through the transformer secondary; so an additional voltage drop represented by the leakage reactance voltage drop of the transformer, must be supplied by the reactor. This tends to be of moderate value because, at the time the transformer primary current passes through zero, the ampere turns produced in the two sections of the secondary by the SR current are equal and opposite so that only the leakage reactance is involved.

Trolley Voltage Requirements

Ignitron rectifiers require pulsed firing voltages to start tube conduction at the beginning of each half-cycle. The firing circuits which provide this voltage are supplied from the same source as the power to the rectifier tubes. When the trolley circuit voltage falls so low that the firing circuits will not generate the pulse, or the
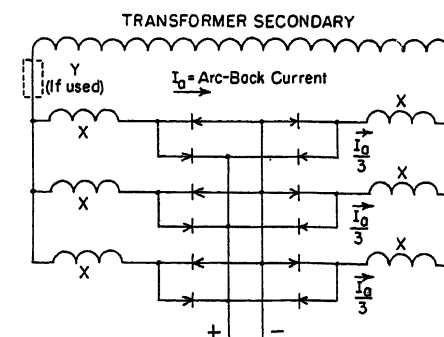
Fig. 3. Comparison of lumped and distributed current-limiting reactance

pulse is delayed in occurring, the main rectifier tubes will not conduct current properly or may fail to conduct entirely. This means that a locomotive or car using ignitron rectifiers will not operate when the trolley voltage falls below the value for which the firing circuits have been designed.

Because extraordinary conditions may result in low voltage on the trolley wire, it appears essential that the firing circuit be designed to provide for locomotive operation over a relatively wide range of voltage. Firing circuit operation is complicated by the fact that rectifier commutation involves a system short circuit during the commutation period and a commutating notch is cut into the applied voltage wave as supplied to all rectifiers on the line at any given time. Firing circuits must be adequate to take a sine wave of voltage with a substantial commutating notch cut out of it and still operate satisfactorily. The commutating notch becomes worse as the supply voltage falls, and also with the increase in trolley circuit reactance as the locomotive traverses the line.

To ensure that the rectifiers will continue to operate under all reasonable low-voltage conditions, voltage regulating equipment may be required for the firing circuits. The firing circuits themselves have an inherent usable voltage range so that a relatively simple voltage regulator with coarse regulation will extend the operating voltage range of the rectifier to acceptable limits.

Rectifier Arc Backs

Probably the most frequent fault to occur on a rectifier locomotive is that of a rectifier tube arcing back. This phenomenon is associated with the tube failing to seal off and insulate at the end of the conducting period. When arcing back, the tube carries current in the reverse direction and effectively short-circuits the secondary of the transformer. Ignitron tubes have a definite limit to the amount of arc-back current they will take for a period of 1 or 2 cycles without damage. These limits are known for most tubes, and where not known they can readily be determined.

The designer has two alternative schemes to limit arc-back currents to safe values.

1. Build sufficient current limiting reactance into the main transformer.

2. Provide separate current limiting reactances for the ignitron tubes.

The first scheme will, in many cases,

TIF OF VARIOUS SINGLE FREQUENCIES					
Frequency	TIF	Frequency	TIF	Frequency	TIF
60	1	900	7270	1740	3660
180	15	1020	11600	1800	3580
300	205	1080	11980	1860	3570
360	370	1140	11100	1980	3500
420	590	1260	7920	2100	3500
540	1250	1380	5470	2220	3530
660	2250	1440	4740	2340	3580
720	2990	1500	4400	3000	3940
780	4080	1620	3900	5000	480

FREQUENCY WEIGHTING CHARACTERISTIC FOR TIF MEASUREMENTS

Fig. 4. Telephone influence weighting factor curve, 1935

mean a larger and heavier transformer than if it were built to its optimum reactance for weight and size. Moreover, high reactance in any part of the rectifier supply circuit penalizes both rectifier and locomotive performance.

Where the number of rectifier tubes permits, the tube circuits may be grouped and the various groups fed through individual current-limiting reactors. The most likely fault is that of a single tube arcing back at any one instant, and sufficient reactance must be provided to limit the current to a safe value for this one tube. Feeding the rectifiers in groups assures equal current division between the groups when all reactances are equal. The greater the number of branch circuits fed through individual reactors, the greater will be the over-all saving in equivalent volt-ampere capacity of these reactors as compared to a single reactor for limiting the current to all rectifier tubes. The extremely high values of current involved in arc backs would quickly saturate any iron that normally would be put in the circuit so best results can be obtained from simple air core reactors. Fig. 3 with the aid of the following shows the operating advantage of distributing the reactance elements of the current-limiting reactor as compared with a lump reactance either in the main transformer or as a separate device.

VOLTAGE LOSS COMPARISON: LUMPED OR DISTRIBUTED REACTANCE FOR CURRENT LIMITING

Let X = required reactance per circuit, ohms. Then for any given tube arcing back, effective resistance in circuit is

$$Y = X + \frac{X}{3} \tag{1}$$

Y is then also the required value of lumped reactance.

Running voltage loss per bridge circuit for distributed X is

$$E_{loss} = 2I_m X \tag{2}$$

and for lumped Y

$$E_{loss} = 3I_m Y = 3I_m\left(X + \frac{X}{3}\right) = 4I_m X \tag{3}$$

Thus, lumped reactance, whether in the transformer or external, gives twice the running loss.

Equation 2 versus equation 3, shows that distributing the reactances in three-bridge-connected rectifier circuits of the 12-tube rectifier selected for this study gives one-half the voltage loss obtained with a lumped reactance. Such a system will therefore give a higher locomotive speed for a given input voltage.

Telephone Interference

The problem of telephone interference may be one of the major considerations in the design of a rectifier car or locomotive. The primary current is far from a sine wave and approaches the trapezoidal wave which would have been obtained had the output of the rectifier been a perfectly flat wave. The primary voltage wave does not stay at zero during the commutating period; instead it assumes a value sufficient to force the short-circuit current through the transformer the current-limiting reactor and the short-circuited rectifier tubes. This produces a nick or depression in the sine wave of primary voltage during commutation. The voltage returns to its normal sine wave value the instant commutation is completed. With a square current wave all odd harmonics, i.e., the third, fifth

PERFORMANCE LOAD POINT 1300 AMPS, 858 VOLTS – MOTOR
TIF (METER READING) – 20.8
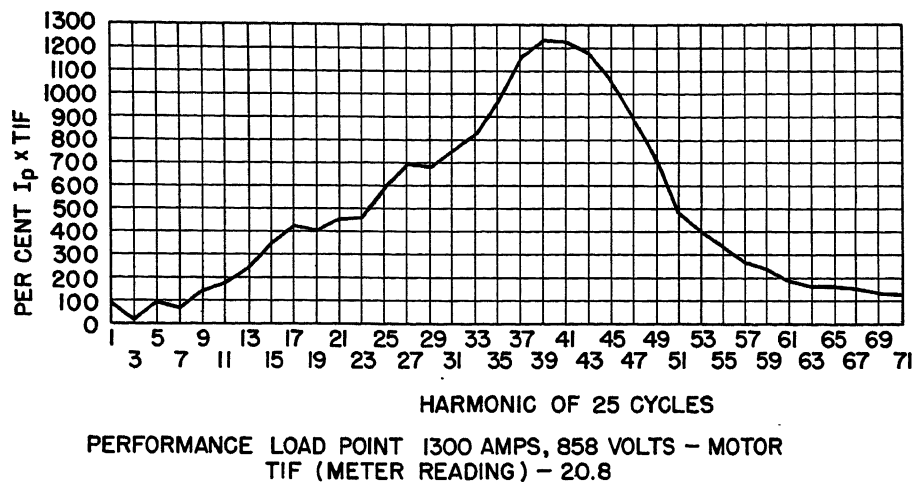
Fig. 5. Weighted harmonic analysis curve of rectifier locomotive without corrective measures

seventh, ninth, etc., occur with decreasing value as the order of the harmonic increases. These harmonics are usually expressed as a percentage of the fundamental. The magnitudes of individual harmonics may be quite significant until the frequency involved is well above the range which would affect telephone communication circuits. It is well known that each of the different frequencies has a different effect on telephone systems and voice sets. The influence of these various frequencies is shown by the 1935 telephone influence weighting factor (TIF) curve in Fig. 4.

A wave analyzer analysis of the primary current wave may be made from which the percentage of the various harmonics may be combined with the TIF curve to give the actual influence of any particular frequency present in the basic current wave. Individual values obtained are the product of the per cent of the harmonic times the weighting factor. The result of such a calculation for a locomotive without corrective measures is shown in Fig. 5. The curves show the peak of the influence curves at the 39th harmonic or approximately 1,000 cycles. Even harmonics are zero and are not shown. The over-all TIF for the current wave is obtained by calculating the root-mean-square or rms value of the area under the curve. This is rather tedious and it is more convenient to utilize a Western Electric 2B noisemeter which measures the noise frequencies in decibels. With the use of a suitable conversion factor the actual value of TIF is readily obtained. For Fig. 5, the TIF was 20.8 as obtained from the meter.

The method described of evaluating TIF is convenient for determining benefits derived from various means of decreasing over-all influence. The TIF can be reduced in a number of ways such as

the following:

1. Connecting a resistance-capacitance circuit across the input to the rectifier tubes.
2. Using the second harmonic filter circuit across the traction motors and main SR or across portions of the latter. See Fig. 1.
3. Phase-shifting two locomotive units or phase-shifting the individual circuits on one locomotive unit.

Since test and experience have shown that equation 2 is not effective in obtaining a significant reduction in TIF, it will not be discussed. Phase-shifting locomotive units (equation 3) has been field-tested by a well-known maker of electric locomotives and a certain amount of success achieved.

On a 3-circuit locomotive unit where the motors are arranged in three groups and the individual groups are fed from their respective bridge-connected rectifiers, phase-shifting of these individual

circuits is possible. The theoretical results of such a 3-circuit phase shift are indicated in Fig. 6. Two curves are shown: the first is the weighted harmonic curve from test data, and the second is the theoretical per-cent remainder curve. A tabulation gives the calculated resultant TIF for several conditions. The minimum value occurs when the per-cent remainder is adjusted to zero at a harmonic somewhat higher than the peak of the weighted harmonic curve at the 46th harmonic. The tabulation also indicates that the minimum does not vary appreciably, even though there is an unbalance in the shift. Such an arrangement gives a calculated improvement in TIF from 39.3 to 10.35.

The curve showing per-cent remainder was derived from the following equation for the shift of the three circuits.

$$X = \frac{\cos N\phi}{3} + \frac{\cos N(\phi+\alpha)}{3} + \frac{\cos N(\phi+2\alpha)}{3}$$

(4)

where

X = remainder ratio due to phase shift
ϕ = instantaneous phase angle, taken as zero
N = harmonic number
α = phase shift of fundamental of each circuit
$\cos N\phi/3$ = contribution of unshifted circuit
$\cos N(\phi+\alpha)/3$ = contribution of circuit shifted α degrees
$\cos N(\phi+2\alpha)/3$ = contribution of circuit shifted 2α degrees

When specified condition is $X = 0$ and $N = 46/2$ the three terms of equation 1 must add to zero and thus they must be 120 degrees apart, and then

ARRANGEMENT	PH. SHIFT	CALC. T.I.F.
SINGLE CIRCUIT	——	39.3
THREE CIRCUIT	5.21°	10.35
THREE CIRCUIT	5.85°	12.1
THREE CIRCUIT	5.21-5.73°	11.5
(10% UNBALANCE)		

PERFORMANCE LOAD POINT 1300 AMPS, 858 VOLTS – MOTOR
3-CIRCUIT PHASE SHIFT PERFORMANCE CURVE ADDED
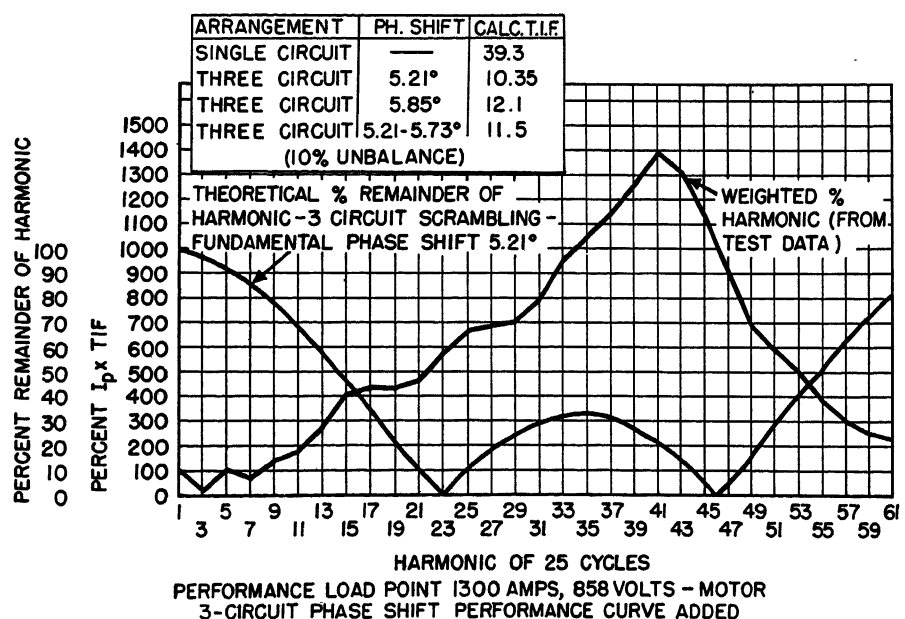
Fig. 6 Weighted harmonic analysis curve of rectifier locomotive with phase-shifting line reactor in supply circuit

PERFORMANCE LOAD POINT 1300 AMPS, 858 VOLTS — MOTOR
TIF (METER READING) — 5.64 RC FILTER — 346.9 mfd
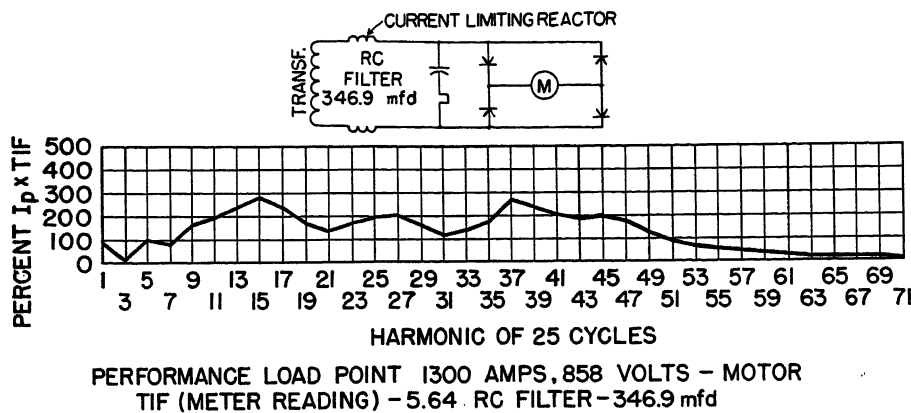
Fig. 7. Weighted harmonic analysis curve of rectifier locomotive with R/C filter in circuit

$$= \frac{2 \times 120}{46} = 5.21 \text{ degrees} \qquad (5)$$

A similar, and generally better, improvement as regards telephone influence can be obtained by the method of item 1 in the section "Telephone Interference." A resistance-capacitance (R/C) circuit is connected across the input to the rectifier tubes on the low side of the current-limiting reactor coils. Fig. 7 shows the connections and the weighted results of this arrangement where the TIF by meter came to 5.64.

Since the addition of the R/C circuit to the rectifier is a much simpler method of improving telephone interference and can be varied over rather wide limits, depending upon the amount of capacitance and resistance added, this scheme is preferred. Also, the addition of the capacitance tends to improve the locomotive power factor, whereas phase-splitting by the addition of reactors makes the power factor worse and at the same time increases the commutating reactance

loss in the circuits in which the reactance is added.

It is a simple matter to simulate telephone interference problems which might be encountered in the field when a stand test is available for testing rectifier equipment. In such a test, an air core reactor in series with the primary of the transformer will provide a magnetic field in which a telephone pickup coil can be located. By its relative location with respect to the series reactor, it will provide any desired degree of noise in the telephone circuit. When the axis of such a pickup coil coincides with the axis of the series reactor, the wave shape which appears in the telephone circuit will correspond to that shown in Fig. 8. An interesting phenomenon occurs when the coil is changed so that its axis is perpendicular to the axis of the series reactor and is properly spaced. A very sharp spike, Fig. 9, appears on the wave at the point where it peaks at the end of rectifier commutation. Noise readings in

the telephone circuits for this type of pickup indicate that a small amount of capacitance connected across the input to the rectifier tubes will make a large reduction in noise. The noise reduction is much more marked than if the R/C circuit is added when the pickup coil is located on the axis of the line reactor. Fig. 10 shows the characteristics of the telephone circuit wave with the pickup coil perpendicular to the axis of the line reactor after the R/C circuit has been added to the rectifier circuit. The sharp spike on the telephone pickup voltage has been completely eliminated. An interesting corollary is that the sharp corner in the primary wave shape has been smoothed and a rather pronounced low-frequency oscillation has developed at the end of the rectifier commutating period.

Practical Aspects of Telephone Interference

A number of field tests have been made for determining the amount of interference which rectifier locomotives generate in adjacent telephone circuits. The results, which have been made available to the technical press, indicate that telephone interference generated by such equipment is much less than has been anticipated.

Tests made on other railroads also substantiate the fact that where telephone lines are run in cable, the noise pickup from rectifier motive power is of minor importance.

On open-wire telephone lines that parallel trolley circuits, the noise pickup is much more pronounced. However, series a-c motor noise is also a problem
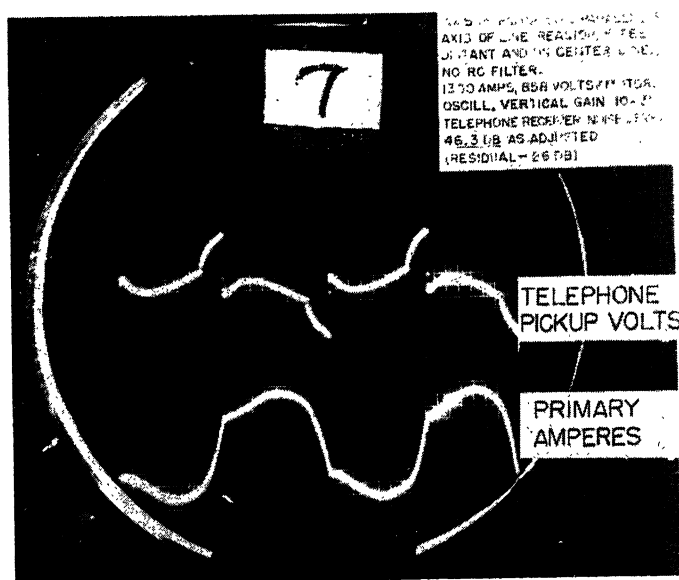
Fig. 8. Wave shape in telephone line without R/C filter
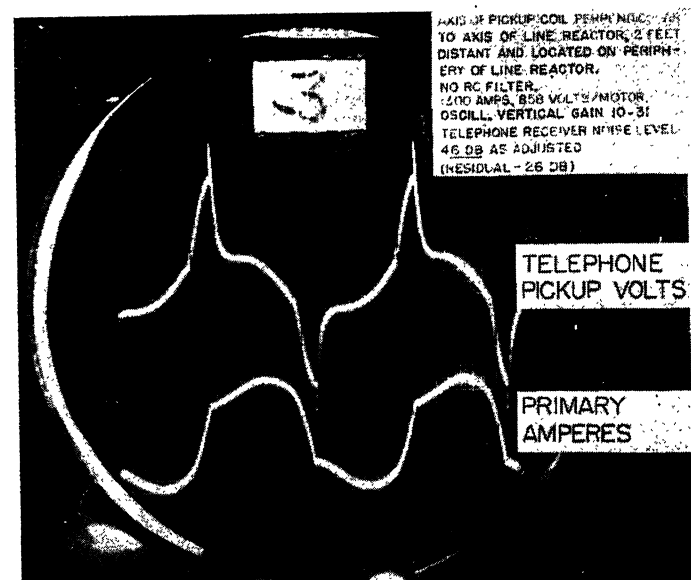
Fig. 9. Wave shape in telephone line with sharp spike at end of rectifier commutation period

1. The traction motor should be a standard low-cost unit developed for Diesel-electric locomotives.

2. D-c smoothing for the benefit of the traction motor is best achieved by an iron core SR built large enough to take care of all requirements.

3. Rectifier tube circuit connections should be of the bridge type rather than the diametric type.

4. Firing circuits should be arranged for satisfactory operation over the maximum trolley voltage variations that will be encountered in service.

5. Main transformers should be built with optimum reactance for minimum weight and size. The additional current-limiting reactance should be built into divided current-limiting reactors, division being in accordance with the basic circuits of the rectifier connections.

6. Telephone interference problems are minor in nature and can be corrected by the addition of a relatively small R/C type filter circuit across the rectifier input.

here, and if the telephone lines are improved to take care of these noises, there is also a marked improvement as far as rectifier equipment is concerned. This equipment can be improved by adding just enough R/C filter circuit across the input to the rectifier tubes to remove the very sharp spike which appears in the picked-up wave on the telephone lines. Field tests have shown that the shape of the wave picked up in the telephone line corresponds very closely to that picked up on the test floor when the pickup coil

is oriented so that its axis is at 90 degrees to the axis of line reactor. As on test, only a small R/C circuit is required to eliminate the spike and make a substantial reduction in the noise level.

Conclusions

The design features which should be incorporated in a rectifier locomotive to ensure maximum performance with minimum weight and cost may be summarized as follows:

References

1. RECTIFIER-TYPE MOTIVE POWER FOR RAILROAD ELECTRIFICATION, L. J. Hibbard, C. C. Whittaker, E. W. Ames. AIEE Transactions, vol. 69, pt. I, 1950, pp. 519–524.

2. THE PENNSYLVANIA RAILROAD IGNITRON-RECTIFIER LOCOMOTIVE, C. C. Whittaker, W. H. Hutchison. AIEE Transactions, vol. 71, pt. II, Jan. 1952, pp. 37–47.

3. RECTIFIER MOTIVE POWER—INDUCTIVE CO-ORDINATION CONSIDERATIONS, E. B. King, K. H. Gordon, L. J. Hibbard. AIEE Transactions, vol. 73, pt. II, July 1954, pp. 107–18.

Discussion

E. B. King (American Telephone and Telegraph Company, New York, N. Y.): In this paper wave shape and telephone interference are given prominence among the considerations in the development of a high-power rectifier locomotive. Of particular interest is the unique method used in simulating telephone interference problems in the laboratory and in determining the amount of improvement afforded by various remedial measures.

I was unable to reconcile, however, the results given in Figs. 5 and 6. In Fig. 5, an over-all current TIF reading of 20.8 was measured, whereas in Fig. 6 a similar curve with substantially the same area under it gives a calculated TIF of 39.3, or about twice as much. The discrepancy seems to be in connection with Fig. 5, as the root-sum-square of its individual harmonics comes out approximately 39. The results of hundreds of tests with the 2B noise measuring set have generally shown quite close agreement of the over-all weighted values read on the meter, when compared with the root-sum-square of the individual weighted harmonics, unless a large fluctuation in the magnitude of some of the harmonics is present during the tests.

Referring to Fig. 4 on the frequency weighting characteristic for TIF measurement, I wish we did have a 1953 weighting

curve as was inadvertently indicated for 1935. It is true that the telephone receiver response before 1935 did peak quite sharply around 1,000 cycles as indicated on the curve, but for many years most of the telephone sets have had receivers whose response characteristics are fairly flat over the voice frequency range. In 1941 a new telephone line weighting curve was adopted, but the proposed 1941 TIF weighting curve was still under discussion at the outbreak of World War II when practically all joint committee work stopped. It is hoped that early consideration of a suitable TIF weighting curve can be given by the interested manufacturers, power companies, and communication companies.

In the section "Practical Aspects of Telephone Interference," some of the statements do not seem to agree with the test results included in unpublished work by Hibbard, Garry, and Loomis in 1955, and the results in the King, Gordon, Hibbard paper (reference 3 of the paper) of 1954. For example, in Fig. 5 of the Hibbard, Garry, Loomis paper, the right-hand curve shows the noise on a telephone cable circuit due to a 6-car rectifier train with no filter, having a noise influence (IT product) of about 4,000. The effect of the train increases the noise on the cable circuit from a quiet 16 db to a maximum of 28 db or a 4-to-1 increase in the noise. Since noise on neighboring communication

circuits is approximately proportional to the noise influence of the motive power, what may be expected when the rectifier locomotive of Fig. 6 runs through the same exposure section? If the load current is 1,000 amperes, the IT product would be approximately 39,000. This is an increase of nearly 10 to 1 or 20 db over the value for the 6-car multiple-unit (MU) rectifier train, and the resulting telephone circuit noise would likely be objectionable. Even with the use of the R/C filter shown in Fig. 7, the IT product for a 1,000-ampere load would be about 5,600, or somewhat higher than for the 6-car MU train without filters.

L. J. Hibbard (Westinghouse Electric Company, East Pittsburgh, Pa.): The papers on rectifier motive power units are a timely addition to the existing published data. It is our considered judgment that rectifier motive power now available is the key to a much wider use of railroad electrification and that the time has come to promote this superior tool.

If rectifier motive power and railroad electrification can now take over its rightful economic place in the railroad transportation field, we are sure that an incentive will be created that will lead to further far-reaching improvements in railroad operation. It will put the railroads in a position to benefit directly from the tremendous

progress being made in the power generation field, and to benefit in the future from the intensive widespread activity in the semiconductor field. It will also be mutually beneficial to many other groups. It will benefit the United States in that it can appreciably steepen the slope of the national electric power consumption curve. Approximately 97 billion kw-hours were consumed by approximately 42,000,000 residential users of electricity in 1953. One of the existing electrified railroads consume 1 to 1¼ billion kw-hours annually. Hence, one electrified railroad uses as much power as 500,000 residential customers, and the load factor of the railroad customer is far better than the load factor of the residential load.

We concur that in the present state of the art the traction motor should be the standard low-cost unit developed for the Diesel-electric locomotive, and that for the present d-c smoothing for the traction motor is best achieved by an iron core reactor. However, when the use of railroad electrification with rectifier motive power has expanded sufficiently to justify the use of a new type of motor, we predict that a pulsating d-c motor will be developed for rectifier-type motive power. This motor will have a laminated stator and will have commutating ability to handle up to at least a 100-per-cent ripple in its propulsion current. The inherent inductance of the motor windings themselves will permit us to omit the d-c choke. Also the decreased inductance in the d-c circuit will further reduce the present minor telephone interference problem substantially.

The benefits obtainable from the use of a pulsating d-c motor on rectifier motive power units were considered 43 years ago. I have in my files a note from B. G. Lamme (former Westinghouse chief engineer) to C. E. Wilson (present United States Secretary of Defense) asking him to design such a motor. The motor was designed, built, and tested by Mr. Wilson. Now that rectifier motive power has successfully been introduced, its development should be of widespread importance.

H. F. Brown (Consulting Engineer, New Haven, Conn.): The authors' several papers on a subject of great interest and importance are most timely.

Several points in this paper seem to require a little further explanation. The efficiency of this R/C circuit for the mitigation of harmonics on rectifier-type motive power has been amply demonstrated in the past by its application to the Pennsylvania Railroad rectifier locomotives and to the New Haven Railroad motor cars. It was natural that it should be applied to these new locomotives in spite of the higher voltage of the circuit. One question arises in this connection: Would not such a circuit be more effective applied directly to the terminals of the transformer secondary, since the prime function is to prevent as far as possible the harmonics passing through the secondary into the primary and the line. The effect of the SR would seem to be the same in either connection.

In conclusion 6 of the paper the author states that telephone interference problems are minor in nature, but from the number

of papers which have been presented on this subject this seems something of an understatement. However, it is significant, and gratifying to know, that the solution of this problem appears to be relatively simple. In connection with past equipment of this type, the amount of such filtering apparatus was determined by road tests. The question arises whether a definite amount of resistance and capacity is designed into this locomotive, or will this also be determined by tests?

Conclusion 3 should not be allowed to stand unchallenged, since all cars and locomotives of this type now in successful operation are of the diametric type of tube circuit connection. The analyses of the two types of connection shown in Fig. 2 do show some advantages for the bridge type of connection, but the disadvantages are entirely omitted.

Two disadvantages, not mentioned, are the necessity for high-voltage switching of the secondary taps and in my opinion, the much more serious one of having two motors, not mechanically coupled, in series. This type of motor connection is not desirable for either high-speed passenger locomotives or for slow-speed freight locomotives. In either case, wheel slipping can occur with attendant motor flashovers and possible tube arc backs. This has always been an inherent disadvantage of some 3,000-volt d-c locomotives, where special slip relays have been installed as a protective measure. Some of the minor disadvantages of the diametric connection would be more than offset by being able to have all the motors always in parallel. This has apparently been the opinion of previous designers of this type of equipment.

Since high-voltage switching had to be used on the secondary taps, it would be of interest to know whether tap-changing on the primary was investigated. This has been used successfully abroad in a number of single-phase locomotive designs, and would appear to have important possibilities with both diametric connections, with the use of multisecondary windings, or the bridge-type connection with its higher voltage.

It is recognized that this locomotive had to be designed with certain axle-loading restrictions, which perhaps limited the transformer weight, and higher voltage possibly meant a lighter secondary winding. On other applications this weight restriction might not be an important factor in the design.

The importance of a wide range of firing circuit voltage to meet the low-voltage conditions of the system cannot be given too much emphasis. Most traction systems are designed with a regulation of about 20 per cent. A-c systems are usually better in this respect than d-c systems. However, emergency conditions often arise on all systems. A transformer station may be out of service, and any locomotive or car, whether of the rectifier type or otherwise, should be able to operate, with all of its important auxiliaries, under such conditions, but of course at slower speed.

The performance of these locomotives in service, which has not yet become effective, will be watched with keen interest, both in this country and abroad.

H. S. Ogden: The question raised by Mr. King on the lack of agreement between *2B* noise measuring set readings and rootsum-square of the individual harmonics as displayed in Figs. 5 and 6 is well taken. When the calculations for the investigation were made which resulted in the information given in Fig. 6, it was found that the even harmonics had been overlooked. As these are all zero the calculation for rootsum-square value came out higher than the *2B* noise meter reading. This approach made no material difference in the results obtained and calculations were not repeated.

I join with Mr. King in hoping that a more representative weighting curve be made available if modern telephone receivers no longer have the response characteristics of the 1935 curve. Mr. King compares the locomotive with a 6-car MU train operating where there is an exposure with a telephone cable. In fairness, a comparison should be made with the 12-car MU train that the New Haven Railroad sometimes operates. The trolley currents are quite comparable when a 12-car train is compared with this locomotive. Contrary to Mr. King's statement, a 6-car train without filter would have a maximum IT product of 11,700. For a 12-car train the IT product would be 23,400. The later figure may be compared with a calculated IT product for the locomotive as obtained from Fig. 5 multiplied by a 1,000-ampere trolley current, i.e., $20.8 \times 1,000 = 20,800$. The IT products are thus roughly equal. Similarly if two can filters per car are used, the IT product for a 12-car train becomes 6,000 or substantially the same as the 5,600 figure calculated by Mr. King for the locomotive.

I appreciate the many good points made by Mr. Hibbard concerning the broad subject of rectifier motive power. Although the use of a specially built pulsating d-c motor for rectifier applications may seem impractical, his historical references are certainly of timely interest.

Mr. Brown questions the point at which the R/C circuit should be connected, suggesting that it be applied directly to the terminals of the transformer secondary. The purpose of the R/C circuit is to provide a low-impedance path in parallel with supply line for by-passing the harmonic currents. This can best be accomplished by ensuring that the supply line has the maximum impedance available ahead of such a parallel path. Thus the filter should be connected as close to the rectifier as possible, throwing everything into the line impedance that one is able to.

Although, as stated in the paper, telephone interference problems are minor in nature, there is no question but that these problems have loomed large in the minds of those who are responsible for the successful operation of locomotives and cars of the rectifier type. That has been and still is the reason for the large amount of published material on the subject.

In answer to Mr. Brown's question, a definite amount of resistance and capacitance has been designed into the locomotive. The adequacy of this will be checked by test.

The analysis given in Fig. 2 presupposes that it is possible to connect two traction motors in series and proceeds from there.

The paper does not undertake to show the relative merits of series-connected motors versus parallel-connected motors. Where d-c motors are used upon a straight electric road locomotive, series motor connections are always used when starting. The conditions on this locomotive are no worse than on such a locomotive and the considerations covered in the paper were determining.

Tap changing on the primary of the main transformer has often been considered for use on electric locomotives in this country, and has been discarded as being economically unsound. It was given only passing consideration when the design of the locomotive referred to by the paper was first considered.

Considerations in Applying D-C Traction Motors on Rectified Single-Phase Power

M. SIMON
ASSOCIATE MEMBER AIEE

IN PAST years various attempts have been made to power locomotives by 25-cycle single-phase rectified voltage applied to d-c series motors. Early efforts were unsuccessful, mainly because of the faulty operation of the mercury-arc rectifier. Continued research and development, however, have now made the operation of d-c traction motors on mercury-arc rectifiers in locomotive service entirely feasible from a technical standpoint. In recent years, a new feature has been added. The development of the Diesel-electric locomotive has resulted in a standardized, low-cost d-c traction motor suitable for general-purpose road-locomotive applications. Hence, any present-day electric locomotive design employing d-c traction motors can benefit by the use of such standarized motors. An example of this is found in the rectifier locomotives recently built by the General Electric Company for the New York, New Haven and Hartford Railroad wherein the standard d-c traction motor shown in Fig. 1 was applied. The problem of operating these motors on rectified single-phase 25-cycle power resolved itself into the problem of smoothing the undulating current to a sufficient degree so that this motor would not be harmfully affected from either a stress or performance standpoint.

Although rectified alternating current has been used for many years as a source of power for d-c motors, it has largely been derived from multiphase rectifiers

operating on 3-phase systems. The current resulting from such a system is essentially pure direct current with an insignificant amount of ripple. On the other hand, the ripple existing in the output of a rectifier operating on single-phase alternating current is of considerable magnitude and would adversely affect both motor performance and maintenance. Fortunately, this ripple, shown in Fig. 2, can be reduced by filter components of practical size to a point where its adverse effects are limited to a tolerable value. Various effects of ripple current on d-c series motors are discussed herein.

Effects of Ripple Current

EDDY CURRENT

When the solid iron parts of a motor are subjected to an alternating or pulsating flux, electromotive forces (emf) are induced in them. These emf produce eddy currents in the iron that are proportional to the time rate of change of flux density. In accordance with Lenz's law, the direction of the eddy currents is such as to oppose any externally applied magnetomotive force tending to change the flux. This phenomenon gives rise to two undesirable effects in a d-c motor. First, if the eddy currents are of sufficient magnitude, they retard the response of the commutating pole flux which impairs commutation. Second, the eddy currents introduce an iron loss in the motor which reduces its efficiency. The total iron loss occurring in the motor as a result of the eddy currents may be divided into two parts: the loss which occurs in the armature and pole faces caused by the rotation of the armature, and the loss

which occurs in the entire magnetic circuit caused by the pulsating magnetic field. The former occurs in any motor whether operating on pure direct current or not, while the latter is the result of a ripple current flowing in the motor circuit. Eddy-current effects can be reduced to a tolerable value by completely laminating the magnetic structure, or by inserting a smoothing reactor in the circuit to limit the ripple current. There are definite design considerations, however, that must be carefully studied before either of the schemes is accepted.

COMMUTATION

Commutation is one of the most important design considerations in a d-c machine. Satisfactory commutation involves the reversal of current in the armature coils as they are short-circuited by a brush from a maximum value in one direction to the same value in the opposite direction, without causing excessive sparking or localized heating of
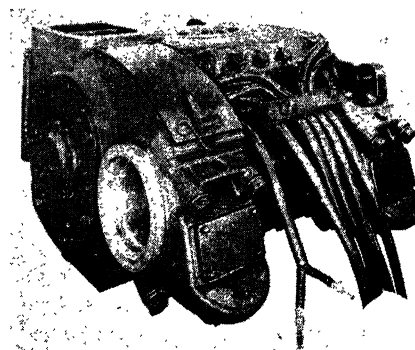
Fig. 1. Standard production heavy-duty type of traction motor, model GE-752-FI
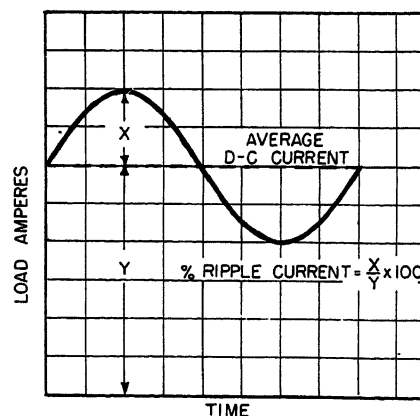
Fig. 2. Definition of per-cent ripple current

the brushes or the commutator. The emf generated in the armature coil by its rotation while short-circuited by the brush opposes commutation if it is generated by the flux of armature reaction. To generate an emf that will produce proper commutation, the commutating field flux must be proportional to, but opposite in direction from, that created by armature reaction.

The addition of a 50-cycle ripple superimposed on the direct current which results from rectifying single-phase 25-cycle alternating current, Fig. 3, can impair commutation. The greatest difficulty will be encountered if the pulsating component of the commutating flux produced by the ripple current is not exactly proportional at every instant to the ripple current flowing in the armature. This condition will result if eddy currents of appreciable magnitude circulate in the solid magnet frame and commutating pole bodies. Under these circumstances, the resulting damped ripple component of the commutating flux will cause the instantaneous flux to be either too great or too small, even though the average value of commutating flux is correct in magnitude. If the ripple current and this angle of lag are too great, the inevitable result will be destructive sparking at the brushes.

Another interesting phenomenon to be taken into consideration when studying the problem of commutation of d-c machines operating on rectified single-phase power is the effect of undulating flux that is present in the main exciting field circuit. In a d-c motor carrying ripple current, an emf is generated in the coil short-circuited under the brush by the variation of this magnetic flux. This emf does not exist when the machine is supplied with pure direct current. If this emf is not reduced, it can cause excessive short-circuit currents which result in destructive commutation. With

the armature coil midway between the exciting field poles, no emf is generated by the rotation of the coil through the magnetic field. The coil, however, encloses the maximum field flux. Since this flux is varying because of the ripple current, the emf it induces by transformer action in the coil is a maximum. This emf, which lags the varying field flux by 90 degrees, is proportional to both the magnitude and frequency of the ripple current present in the exciting field circuit, but is independent of the speed of rotation.

Fortunately the eddy currents induced in the solid magnet frame reduce the magnitude of the varying flux at the main exciting pole air gap to the extent that commutation is not noticeably influenced. It is interesting to note that the eddy currents aid in reducing the main field alternating flux effects, but hinder the phase relationship of commutating flux with armature flux.

As has been briefly explained, the commutation of a d-c motor operating on rectified single-phase power is directly affected by the magnitude of the ripple current present in its circuit. It is, therefore, essential that this current be limited sufficiently to assure acceptable commutation throughout the load range over which the motor is to operate. The obvious means of doing this is by the use of a smoothing reactor in the motor circuit. Before determining the amount of smoothing reactor required from the standpoint of motor commutation, it is necessary to establish the maximum sparking at the brushes that can be tolerated without undesirable commutator pitting and excessive brush wear. This commutation level was determined from cyclic load tests on a standard production, heavy-duty type of d-c traction motor operating on rectified single-phase 25-cycle power.

This test consisted essentially of operating the motor throughout a volt-ampere range simulating expected service load conditions. The duty cycle chosen for this test was the d-c traction motor performance requirement of a rectifier locomotive operating between Grand Central Terminal, New York, and New Haven, Connecticut on the New York, New Haven and Hartford Railroad. The test was set up for unattended continuous operation with only brief interruptions of power for periodic inspection of commutator and brushes. Results after several weeks of operation showed that the motor could operate continuously with sparking at the brushes equal to 6 to 8 sparks per brush without any evidence of commutator film breakdown or commutator pitting.

Once this sparking limit had been established, commutation tests were continued with the motor operating on rectified power to determine the size of smoothing reactor required to give performance comparable to its normal direct voltage application. Similar commutation tests were run with the motor operating on pure direct current. In this case, the sparking was limited to 3 to 4 sparks per brush because of motor design limitations that would be exceeded for the 6 to 8 sparks per brush. The volt-ampere curves of Fig. 3 show the range of operation of this motor on both rectified and d-c power. A curve representing the per-cent ripple current flowing in the motor circuit is also shown in Fig. 3.

It may be of interest here to point out that the aforementioned scheme using a large smoothing reactor with a standardized motor is by no means the only approach to the solution of the problem encountered in operating d-c traction motors on rectified single-phase power.
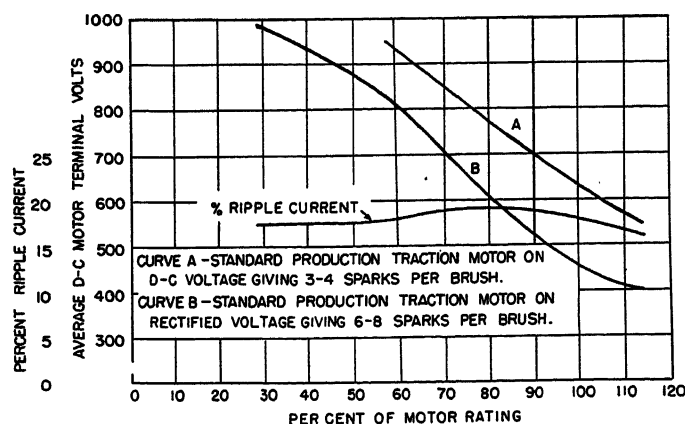
Fig. 3. Commutation comparison curves on direct current and rectified alternating current
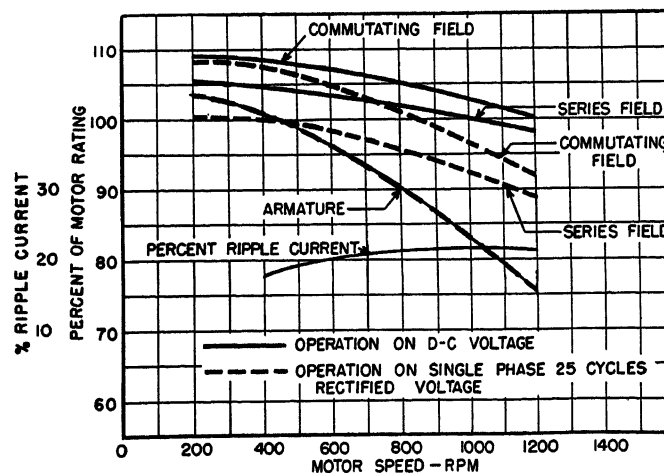
Fig. 4. Continuous rating comparison curves for operation on direct current and rectified alternating current, 100-per-cent field test
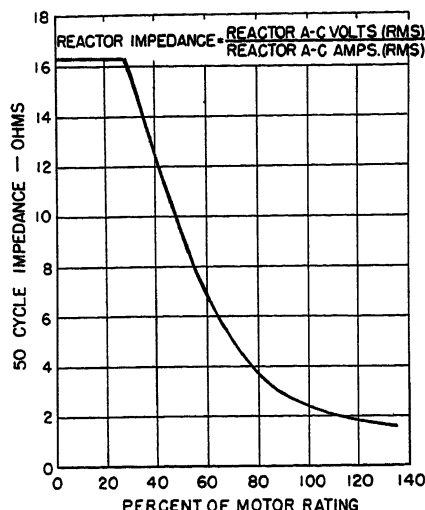
Fig. 5 (left). Smoothing reactor impedance characteristic
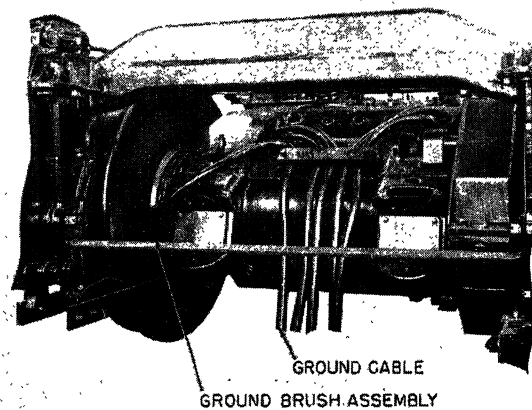
Fig. 6 (right). Truck-mounted traction motor showing ground cable connection to ground brush assembly

GROUND CABLE

GROUND BRUSH ASSEMBLY

This method of approach, however, possesses the advantage of employing a low-cost standard d-c motor, even though the smoothing reactor is of considerable size. Design alterations could be incorporated in the motor to enable it to carry more ripple current, thereby reducing the size of the smoothing reactor. Such a motor would, however, be similar in design and cost to a completely laminated machine and the cost advantages of the standardized motor would be lost.

HEATING

Motor heating is equally as important as commutation and must be carefully considered before applying d-c traction motors on rectified, single-phase 25-cycle power. As previously mentioned, the pulsating flux in the motor is greatly dampened by the eddy currents induced in the solid magnet frame. These eddy currents cause a core loss in the frame which, in turn, increases the operating temperature of the frame and the heating of the field windings. The operating temperature of the laminated armature is not noticeably affected by the presence of the ripple current because by its rotation it is normally subjected to frequencies and magnitudes of armature flux variation in excess of that produced by the 50-cycle ripple current. Numerous continuous heat runs were made on the standard d-c traction motor operating on rectified power. These tests were conducted with the minimum smoothing reactor impedance in the motor circuit as determined from the commutation tests. The results are plotted in Fig. 4. Here each motor winding can be compared with its pure d-c rating for the same temperature rise. A curve representing the percent ripple current is also included in Fig. 4.

TORQUE PULSATIONS

The pulsating current in the motor circuit not only influences the electrical characteristics but also affects the mechanical stresses in the motor. Strain gauge tests indicate that a 50-cycle torque pulsation exists in the motor shaft. With solid gears this torque measured ±9 per cent of steady-state full load torque. With resilient gears it was reduced to ±4 per cent. Although the resilient gear effectively reduced the magnitude of the pulsating torque, the value with standard solid gears is not great enough to cause mechanical failures. Hence, an expensive resilient gear is not justified.

It is interesting to note that tests and calculations show that the first torsional node of vibration for the traction motor is very low—on the order of 8 cycles per second—because of the relatively soft springs in the motor nose suspension. For this reason, the nose suspension functions as a very effective vibration isolator for frequencies higher than 15 cycles per second. The function of the resilient gear is the same—vibration isolation. Its benefit is small, however, when the natural frequency of the system is well below the exciting frequency.

Smoothing Reactor

The design of a smoothing reactor for this motor can be varied depending upon the space and weight limitations. For locomotive application where space and weight are limited, these reactors are kept as small and light as possible. This requires that the reactor be designed around the commutating ability of the motor. For a given motor, the theoretical commutating ability of the motor varies throughout its load range of operation. Motor application for high values of current is generally limited by heating; therefore, the voltage application is

lowered. For low values of current, motor application is limited by voltage (primarily volts per bar) which for this motor is in the vicinity of 1,100 volts. The characteristics of the motor are such that the allowable ripple current (a-c component) is a constant value regardless of the magnitude of the d-c component. This motor characteristic permits the use of a reactor which saturates with load. The resultant design is smaller and lighter than one of the nonsaturating type, with sufficient inductance to limit the ripple current when the motor is operated on maximum application voltage. A reactor of this design has a 50-cycle impedance curve, as shown in Fig. 5. It is of interest to note that the ratio of unsaturated to saturated impedance for this reactor is 8 to 1. This is an added factor in making the reactor small.

An easy way of calculating the effective 50-cycle impedance characteristic of the reactor in the saturated range is to apply the equation

$$Z = 2fN\frac{\phi}{I} \times 10^{-7}$$

where

Z = impedance, ohms
f = frequency, cycles per second
N = total turns in series
ϕ/I = change of flux (lines) with respect to corresponding change in current, amperes

Reactor impedances, as determined from tests by the a-c volt-ammeter method, deviate slightly from the calculated values. This is accounted for by the fact that the ripple current is not a true sine wave.

Motor Modifications

As installed on the rectifier locomotive, the standard d-c traction motor as applied with the rectifier locomotive was slightly modified for 660-volt third-rail

operation into Grand Central Terminal. One of these modifications was provision for a ground return current path to by-pass the 660-volt current around the locomotive roller-bearing journal boxes. The standard axle lining flange dust guard was replaced with a combination dust guard and ground brush assembly. The ground return cable is connected to the ground brush assembly; see Fig. 6.

Another modification consisted of a flash ring assembly designed to provide motor armature bearing protection by directing flashover currents to ground through the ring. These modifications are designed as features which can be easily added to the standard Diesel-electric traction motor.

Conclusions

Because of the relatively small quantity of rectifier locomotives involved, the most economical application of a d-c traction motor on rectified single-phase 25-cycle power is one that employs a standard production-type motor with sufficient series reactors to limit the ripple properly. Motors possessing inherent load characteristics which permit the use of a reactor that saturates with load have definite advantages for such applications. A further development that would bene-fit this type of power transformation includes d-c motor operation on rectified 60-cycle power.

Discussion

L. J. Hibbard (Westinghouse Electric Corporation, East Pittsburgh, Pa.): We have tested five different traction motors on rectified direct current. In each case, we determined by test how low the ripple had to be so as to obtain approximately the same grade of commutation on rectified direct current as on direct current. In each case, our test facilities were such that we could change from direct current to rectified direct current or vice versa merely by throwing a switch.

The permissible ripple varies for different designs of motors. We have one type of motor in operation with approximately a 35-per-cent ripple at continuous rating which gives black commutation throughout its operating range on rectified direct current as well as on direct current. Another type of motor needed approximately 22- to 25-per-cent ripple at continuous rating to make the rectified d-c commutation approximately match d-c commutation. Other types of motors were satisfactory at 28- to 30-per-cent ripple (ripple at continuous rating).

In all of the foregoing applications, the d-c choke characteristics are such that the per-cent ripple decreases at heavier current and increases at lighter currents. In each application we expect the commutator and brush life on rectified direct current to match fully the commutator and brush life obtained on the same motor in d-c service.

In fact, we expect better motor life on rectifier motive power with all motors in parallel operating on a flat speed-current characteristic than can be obtained on the same motor in d-c operation with two or more motors in series on a steep speed-T.E. characteristic, e.g., Diesel-electric locomotives.

It is also interesting to note that, when the ripple value is low enough to give us the commutation characteristics desired, we were unable to measure any winding temperature differences between rectified d-c and d-c operation. This is true for the main field and commutating field as well as for the armature winding.

Two 8-hour tests were made in 1949 on Pennsylvania Railroad car no. 4561 in local service between Broad Street Station and Paoli and return. One 8-hour test was made with the d-c choke adjusted to give approximately 35-per-cent ripple at the continuous rating point with a lower ripple at heavier currents and a higher ripple at lighter currents. Thermocouples were applied to all traction motor field windings, chokes, and other main circuit apparatus and recorded with a 16-element Brown temperature recorder.

The next day an 8-hour test was made under exactly the same conditions with a d-c filter around the traction motor windings so that the ripple was reduced to 5 per cent or less in the motor windings. We obtained the same temperature readings for both tests.

H. F. Brown (Gibbs and Hill, Inc., New York, N. Y.): Mr. Simon has presented an excellent addition to the engineering literature. Mention is made of the impairment in efficiency of the motor because of the effects of the undulating current. It would be of interest to have the author show a comparison between the efficiency of the motor as used on direct current on a Diesel-electric locomotive and on the 25-cycle rectifier locomotive.

Further information on this subject is to be found in reference 1. A most significant statement in the paper occurs in the last sentence in which 60-cycle railway electrification with rectifier equipment is foreseen and endorsed.

REFERENCE

1. M. Blondet. French National Railways, Annency, France, 1951.

A New Power Supply for Railway Cars

E. F. BREDENBERG
ASSOCIATE MEMBER AIEE

THE usual d-c power supply for railway cars consists basically of a battery to which the car load is connected, a d-c generator, and a means of controlling the generator and connecting it to the load. During the last 15 years, a successful system has been established in the general range of 20 to 40 kw. The rotating equipment is a motor-generator set consisting of a d-c generator, a 3-phase 60-cycle 220-volt induction motor, and an armature reversing switch. While the car is in motion, this set is driven from a car axle through a drive unit and clutch. While the car is standing, the set is driven from wayside power through its induction motor.

Exacting specifications as well as the necessity for extreme simplicity and reduced maintenance indicated the desirability of a new type of power supply. This need has been met with a motor-generator-exciter system which with its associated control represents a major advance in the art of railway-car power-supply design.

System Specifications

At first thought, it may seem relatively simple to operate a d-c generator connected to a battery and its load. Railway service, however, presents such a multiplicity of requirements and limitations that the task is really formidable.

The system must be automatic, reliable and must require a minimum amount of maintenance. Often an axle generator is called upon to operate successfully over a wide range of temperatures and atmospheric conditions. A north-south train, for instance, frequently passes from extreme snow and cold to summer temperatures in a single run. The rotating equipment must therefore be designed to
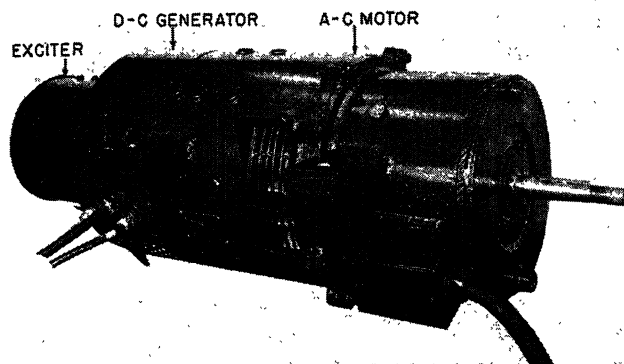
Fig. 1 (above). Side view of axle-driven motor-generator set
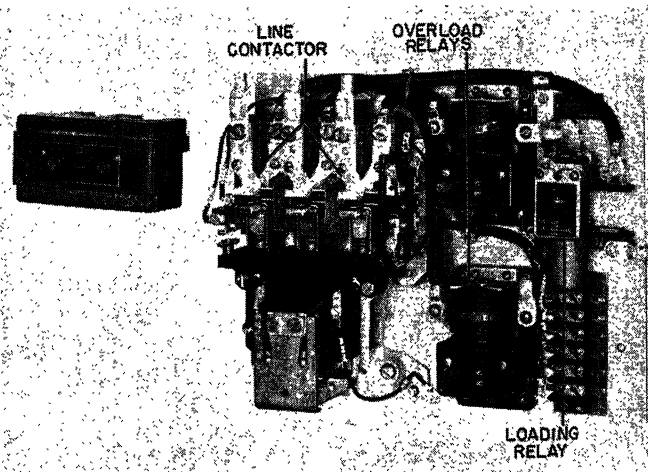
Fig. 2 (right). A-c control panel with cover removed

keep out snow and water, and yet not overheat when operating in high ambient temperatures.

The equipment must provide satisfactory supply to the connected load at essentially constant voltage over a wide speed range. The speeds, for operation from 85-per-cent voltage at no load to full voltage at full load, vary from approximately 5 to 1 to as great as 7 to 1. Full load and full voltage must be held over most of the range. Control must be supplied to assure that a limit is placed on the maximum current which can be taken from the generator. This is necessary for the protection of both the rotating equipment and the battery.

The generator and regulator must be co-ordinated to provide stable operation. At any speed throughout the entire range, the generator must be able to accept the instant application and dumping of full load and, at the same time, maintain this stable operation. It is commonly known that a d-c machine is unstable if, at any speed, its excitation requirements decrease as the load increases. As the speed range increases, the problem of designing a stable machine becomes more difficult. The system here described presented an added problem in that both the generator and exciter had to be designed as stable machines. Also the cumulative effect of the two potential instabilities on the system output had to be eliminated.

Because of the different capacities of batteries used by various railroads, an axle generator design must provide for nominal systems of 40, 80, and 140 volts. An adequate power supply must be produced for use with either nickel-alkaline or lead-acid batteries at each of these general voltages. The control must connect the generator to the battery only when the generator voltage is slightly higher than the battery voltage and must disconnect it when current flows from the

battery to the generator. It is extremely important that the differential voltage be small when the generator and battery are brought together to avoid excessive flicker of the car lights.

Former Equipment and Limitations

Before design work on a new system could be started, it was necessary to make an extensive investigation to determine the limitations of existing equipment and the improvements needed to provide complete customer satisfaction.

In former equipment design a voltage regulator was employed which acted directly on the generator field. This was a multiple-fingered relay operated by coils through "motor action." The relay fingers commutate resistance in series with the generator field. Both voltage-limit and current-limit features are provided on the same structure, although the

Fig. 3. D-c control panel with cover removed

current-limit coil receives its signal from a shunt connected in the load circuit.

Customers have recognized that the performance of this regulator is excellent. However, general opinion has been that a simplified regulating system would be more easily and economically maintained. In addition, the wide speed ranges make it difficult to handle the necessary volt-amperes of regulator finger duty, especially for the 140-volt system. This pointed to an amplified excitation system which would permit the use of a simple, inexpensive regulating relay.

It was also found advisable to design a system with inherent characteristics that eliminated such problems as low batteries and reverse polarity. Other improvements found to be welcome were a drastic reduction in pressing tonnage for the generator shaft and the elimination of the armature reversing switch. In short, the data accumulated showed a need for an over-all simplified equipment built especially for the maintenance man.

New Equipment

Since the a-c control for wayside power operation presently in use was completely adequate in its protection of the a-c motor and the a-c power supply, it was not changed. The following, therefore, is concerned mainly with the d-c portion of the new equipment.

COMPONENTS

The new rotating equipment, Fig. 1, consists of a 4-pole d-c shunt generator, a 6-pole exciter, and a 4-pole 3-phase 60-cycle 220-volt induction motor assembled on a common shaft. The generator armature, fan, and a-c rotor are on a common sleeve to provide for a minimum pressing tonnage. This machine, as in former equipments, is driven either from the car axle through a drive unit and

Bredenberg—New Power Supply for Railway Cars   JULY 1955
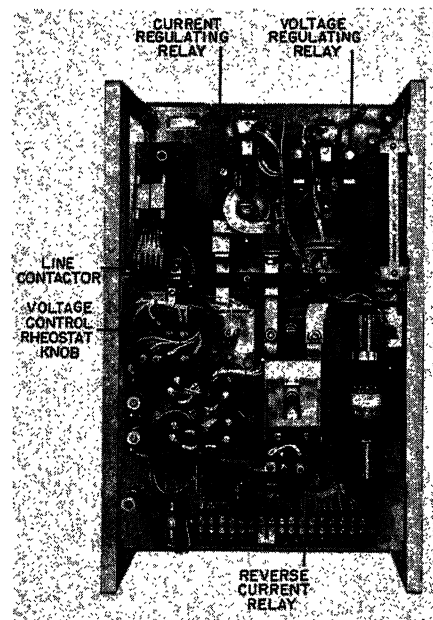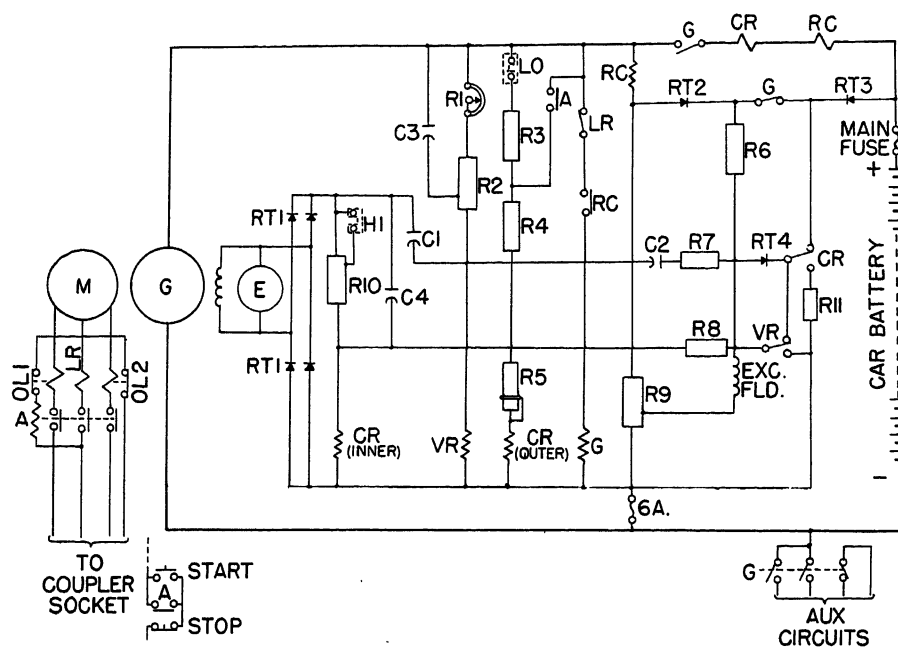
Fig. 4. Schematic connection diagram

clutch, or from wayside power through its induction motor.

The control equipment includes an a-c panel, Fig. 2, mounting an a-c line contactor, two overload relays, and an unloading relay. The d-c portion of the control equipment, Fig. 3, is a panel mounting a line contactor, a voltage-regulating relay, a current-regulating relay, a reverse-current relay, and a fuse or overload relay. Also mounted on a d-c panel are miscellaneous resistors, capacitors, and rectifiers for control and stabilizing circuits. The schematic connections are shown in Fig. 4.

OPERATION

When the generator and exciter are not running, current is fed from the battery through rectifier RT3 to the exciter field. Rectifier RT2 blocks this current so that it will not flow to the generator. As the

set comes up to speed, the exciter builds up and excites the generator field. Because the exciter field polarity is always the same, the exciter armature polarity will reverse with a reversal of rotation of the set. Since the generator field is directly across the exciter armature, the generator field will likewise reverse with set rotation. Since both rotation and field current are reversed, the generator polarity will remain the same. Therefore an armature reversing switch is unnecessary. The control of the reverse-current relay, the voltage-regulating relay, and the current-regulating relay is of major importance.

The reverse-current relay RC, Fig. 5, is a simple 2-coil relay having a series and a shunt coil wound on the same core. A knife-edged, counterbalanced armature carries the moving contacts. The fixed contacts are carried on a finger block

which is mounted on the armature stop. There is one normally open circuit, the only purpose of which is to energize and de-energize the coil of the generator contactor. The only adjustment on the reverse-current relay is a drop-out screw mounted on the armature.

On build-up, when the generator voltage is lower than the battery voltage, the exciter field is supplied only through rectifier RT3. The RC shunt coil sees current only in series with resistor R9. Rectifier RT2 is blocking. When the generator voltage rises slightly above battery voltage, rectifier RT2 begins to conduct and rectifier RT3 begins to block. A considerable increase in current then occurs in the RC shunt coil with only very small increase in generator voltage. The RC relay then picks up energizes the generator contactor G and connects the generator to the battery. The generator voltage then rises to the proper battery-charging voltage where it is regulated by the voltage-regulating relay. An interlock on contactor G opens to provide the reverse-current relay drop-out setting. Exciter field current is then supplied either through RT3 or through RT2 depending on the position of the contacts of the voltage-regulating relay VR and the current-regulating relay CR.

As long as the generator voltage is higher than the battery voltage, the series coil of RC aids the shunt coil. When reverse current flows, the series coil bucks the shunt coil. This causes RC to drop out and open the generator contactor.

The voltage-regulating relay VR, Fig. 6, has a single coil and a nonmagnetic armature. A moveable contact bar which
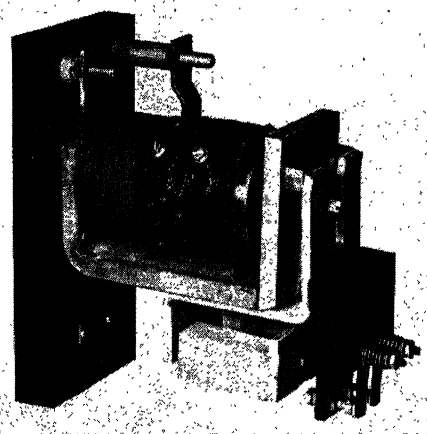
Fig. 5. Reverse-current relay
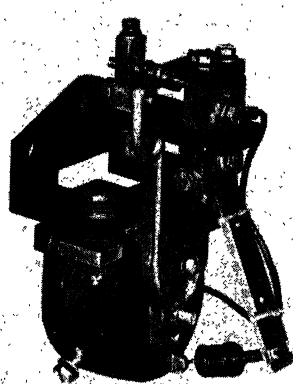
Fig. 6. Voltage-regulating relay
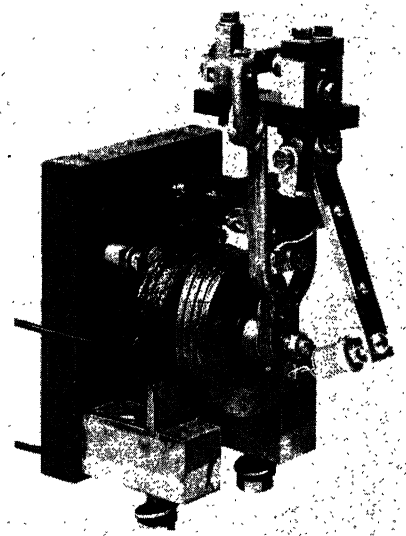
Fig. 7. Current-regulating relay

operates between two fixed contacts is mounted on the armature. Also mounted on the armature is a magnetic core plug that operates in the air gap at the center of the coil. A fixed core plug is mounted at the bottom of the air gap. The armature is held in its normal position by a spring. The adjustments of the movable core plug and the armature spring are sealed. The shunt coil is connected across the generator output in series with a resistor and rheostat. The rheostat is adjusted to set the relay to regulate at the required battery-charging voltage. The relay contacts are connected in the exciter field circuit.

When the generator voltage rises to the value to be regulated, VR starts to pick up. This reduces the exciter field current and causes a corresponding decrease in generator voltage. VR then drops back to its original position. This action continues and the relay vibrates on its normally closed contacts. As the generator speed increases, VR operates more in the open position than in the closed and, finally, the contacts move to the outer side and vibrate against the normally open contact.

The current-regulating relay CR, Fig. 7, is similar in construction to the voltage-regulating relay. It has two bias coils and a coil connected in series with the generator output. Again, the core and spring adjustments are sealed. When the line current reaches a predetermined value, CR picks up and reduces generator voltage. This drops out VR. Then CR takes over the operation of the system and vibrates against its stationary contacts to hold the proper amount of current.

Rectifier $RT1$ provides for a constant polarity supply of exciter output voltage to the CR inner bias coil. As speed increases, exciter voltage decreases and current through the CR bias coil decreases thus increasing the series current required to operate CR. This provides an increased generator current-limit setting to take advantage of increased ratings as speed rises. Similarly, the CR outer bias coil, through an a-c line contactor interlock, adjusts the CR relay to a special wayside power current-limit setting to protect both the a-c motor and the wayside power facilities.

Several stabilizing circuits give satisfactory operation over the complete range of speed and load. These circuits prevent hunting of the regulating relays and subsequent instability in the generator output. In short, when sudden system changes occur, the stabilizing circuits act to oppose the changes. Whether under steady-state or transient conditions, the voltage- and current-regulating relays must be stabilized so that the moving contact will vibrate continuously against the fixed contact.

When the contact bar of relay VR moves, capacitor $C2$ and resistor $R7$ cause an increase or decrease in the current through the VR coil to oppose the original motion. In this manner, VR vibrates continuously as long as it is controlling the system. In like manner, resistor $R8$ causes an increase or decrease in the CR inner bias coil current to keep CR vibrating.

When a sudden rise occurs in the exciter voltage while operating in voltage limit capacitor $C1$ feeds a signal to the VR coil to pick up VR and reduce exciter excitation. During current-limit operation, a sudden rise in the exciter voltage will be opposed by capacitor $C4$ which feeds a signal to the CR inner coil. A sudden rise in generator voltage is opposed by capacitor $C3$, which feeds its signal to the VR coil.

The equipment is designed as far as possible with an eye to the future. Current coils in particular are rated for use with low- and high-capacity generators. High and low settings of current limit are provided so that the same panel may be used with different generators. Devices have been selected to include as many common parts as possible for each of the three voltage ranges.

Advantages

This new design is largely one of circuitry. To a great extent it has been possible to use proved circuit components. Because of the amplification provided by the exciter, the regulating-relay contacts are applied at a volt-ampere level considerably below their rating. Capacitors and rectifiers, in particular, have been chosen conservatively to assure long life. The low level of control power, compared to former systems, materially reduces the temperature of the control locker and adds to the life of all equipment mounted in it.

Since the regulating-relay adjustments are sealed, there is no possibility of tampering and consequent maladjustment of the system. Design is such that individual components can be tested separately and installed without the necessity of a system test. This should materially reduce maintenance problems.

The multiple-finger voltage regulator formerly used has been replaced by two simple relays. These have a total of two double-throw contacts instead of 20 contacts and weigh only about one-sixth as much as the former regulator. Also, the following changes have been made:

The armature reversing switch has been eliminated.

Shaft-pressing tonnages have been greatly reduced.

Available speed ranges have been increased.

These changes have resulted in an equipment that is characterized by simplicity, long life, and minimum maintenance. Operating experience with more than 100 of the new equipments on several major railroads indicate that they are providing reliable and satisfactory service.

◆

Discussion

W. J. Madden (Pennsylvania Railroad Company, Philadelphia, Pa.): With this type of apparatus, there are two general methods of control. In one system, the generator voltage is controlled directly by varying the generator shunt field strength, while the other system employs a separate exciter and the generator voltage is controlled by varying the strength of the exciter field.

The axle generators presently in use employ the former method of control and utilize, in addition to the shunt field, a series commutating field which provides good commutation and stable operation over the wide range of car speeds and loads. With this type of control, the voltage regulator is required to handle a somewhat heavier current than is the case with systems employing a separate exciter. One of the large eastern railroads has approximately 1,200 machines of this self-excited general type, some of which have been in service approximately 20 years.

The method of mounting the armature laminations on the shaft has always represented a problem. If the laminations are assembled on a quill and the quill is mounted on the armature shaft, the quill will occupy much needed space for armature iron. This condition becomes critical on the larger capacity machines because an increase in the outside diameter of the machine to compensate for the loss of iron occupied by the quill would probably encroach on the allowable space between the bottom of the generator and the clearance line at top of rail. Further, there are space limitations if an attempt is made to extend the length of the machine.

Space is at a premium in the electrical control cabinets in passenger cars and it is desirable for simplicity of application and accessibility for maintenance to have the following conditions:

1. Power cable connections arranged at the bottom of the panel.

2. Terminal posts and mounting screws

of component items on the panel readily accessible without the necessity for removing other items.

The addition of the angle frame work to support the cover on the d-c control panel will make it difficult to inspect and maintain the component parts along the edges of this panel as well as along the edges of adjacent panels.

The question of whether to use fuses or circuit breakers for protection is always open for discussion. In passenger car service, fuses have the following objectionable features:

1. There is always the possibility that the wrong capacity fuse will be applied.

2. Frequently, the jaws or clips of the fuse holder are distorted, causing overheating due to poor contact and finally blowing of the fuse.

In the d-c control circuits, it is noted that a large number of resistors are employed and, even though the amperes handled are not heavy, considerable trouble can develop because of overheating if the resistors are not selected with the proper wattage rating. This is particularly true in passenger car work where the ambient temperature of the electric compartment may be appreciably higher than that in the body of the car.

This same condition of wattage rating applies to the operating coils of the several relays.

The use of vibrating-type regulators for control of voltage and current on cabin-car axle-driven alternator-rectifier systems has proved satisfactory. In this service, the regulators function with a wide range of car speeds from 8 to 90 miles per hour. The regulators described by the author are similar to those used on the cabin cars. Based on my experience with the type used on the cabin cars, the similar type proposed for passenger car service should give acceptable performance.

E. F. Bredenberg: I wish to thank Mr. Madden for his interest. There are several comments which require further discussion.

Before producing any of the new equipments, we conducted an extensive investigation to determine the control and rotating equipment features which would be most acceptable to the majority of users. The points in question were carefully considered in this investigation as it is recognized that the various car builders and railroads are not in complete agreement.

Concerning the location of the two power connections to the d-c panel, we have produced, or have on order, approximately 200 of the new equipments. To date, all custo-

mers and car builders have found it desirable to connect the two power cables at the sides of the panel. If the power connections were made at the bottom, more effective panel height would be needed.

The objection to the angle frame work or rails on the d-c panel has been overcome. On the first group of panels, these rails were welded to the panel. We are now producing panels having bolted rails so that the user can remove the rails if desired. We prefer to include rails for the purpose of protection of the panel components during handling and shipping. Again, all users to date have elected to keep the rails as a desirable feature.

As Mr. Madden indicates, the various railroads are divided as to the use of fuses or overload relays in the main generator circuit. We, accordingly, have decided to furnish a fuse as standard equipment, but will supply an overload relay if desired.

Resistors as used in these circuits are applied at 50 per cent or less of their rated wattage to prevent overheating. Also, control-locker temperatures are considerably lower where the new equipment is used because there is no need for a generator field resistor. This provides an even safer control resistor application. In short, all known customer's problems have been investigated to assure a system which is most advantageous to all.

Rectifier Locomotives for the New York, New Haven and Hartford Railroad

F. D. GOWANS
NONMEMBER AIEE

IN DECEMBER 1954, delivery was begun on an order of ten rectifier-type locomotives for the New York, New Haven and Hartford Railroad. These are high-speed passenger locomotives operating over the Railroad's electrified line between New York City and New Haven.

Although quite similar in characteristics and performance to other type of locomotives now being used in this service, the new motive power differs from them in many electrical and mechanical features. To appreciate these it is necessary to understand the unusual nature of the New Haven passenger locomotive operation into New York City. The railroad's main line is electrified with an 11,000-volt 25-cycle single-phase overhead a-c system from New Haven to Pennsylvania Station in New York. From Woodlawn, going into Grand Central Terminal over the Park Avenue viaduct, the line is electrified

with a 660-volt d-c third-rail system. Passenger locomotives must, therefore, be designed for operation from these two power sources; they must carry complete train-heating equipment and supplies, yet must not exceed the axle loading limit of 58,000 pounds over the viaduct.

Application

Until recently, the use of rectifiers in railway applications has been limited largely to substations furnishing d-c power for use on locomotives and cars. The first large domestic application of rectifiers as part of the locomotive or car propulsion equipment began in 1954 when the New Haven Railroad received delivery of 100 multiple-unit cars, equipped by the Westinghouse Electric Corporation, and the first of this locomotive order. Whether installed in substations or on rolling stock, the

rectifier has facilitated the use of d-c traction motors on a-c motive power.

In recent years, the d-c traction motor has benefited from intensive development for the Diesel-electric locomotive. Facilities for quantity production and experience gained from thousands of motors in service have greatly reduced the manufacturing cost and maintenance expense of this type of traction motor. The rectifier locomotives described from this development through the use of General Electric's standard *GE-752* d-c motor.

Description

The locomotive is shown in Fig. 1. The cab of box-type construction with streamlined ends is fabricated from sheets and structural shapes by welding. The cab sides are designed as load-carrying girders which support the underframe and equipment. Each end of the underframe is fitted with a pilot, rubber draft gear, and tight-lock coupler. New Haven Railroad Standard head-
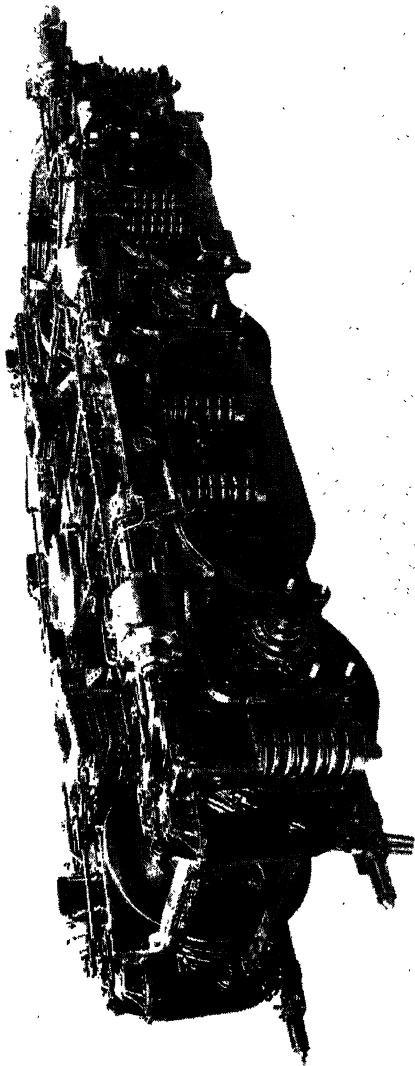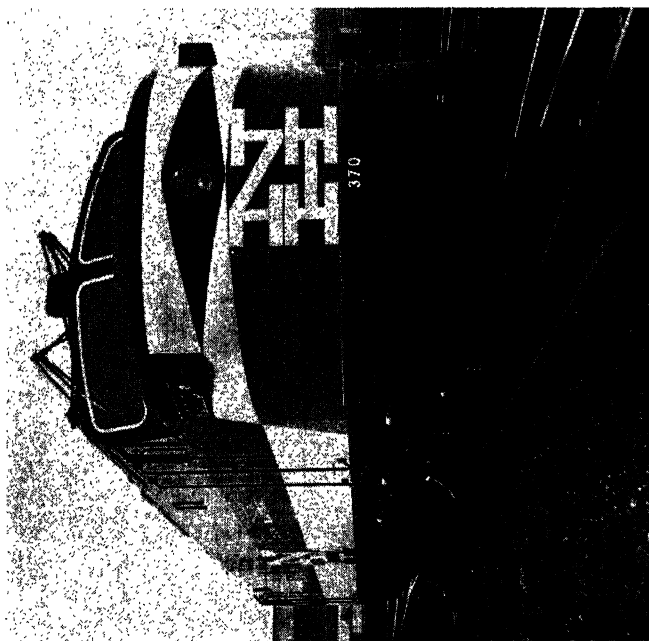
Fig. 1 (left). Appearance of completed rectifier locomotive

Fig. 2 (above). Three-quarter view of truck with motors assembled

Fig. 3 (below). Location of principal pieces of apparatus in locomotive cab

Gowans—Rectifier Locomotives for the New Haven Railroad
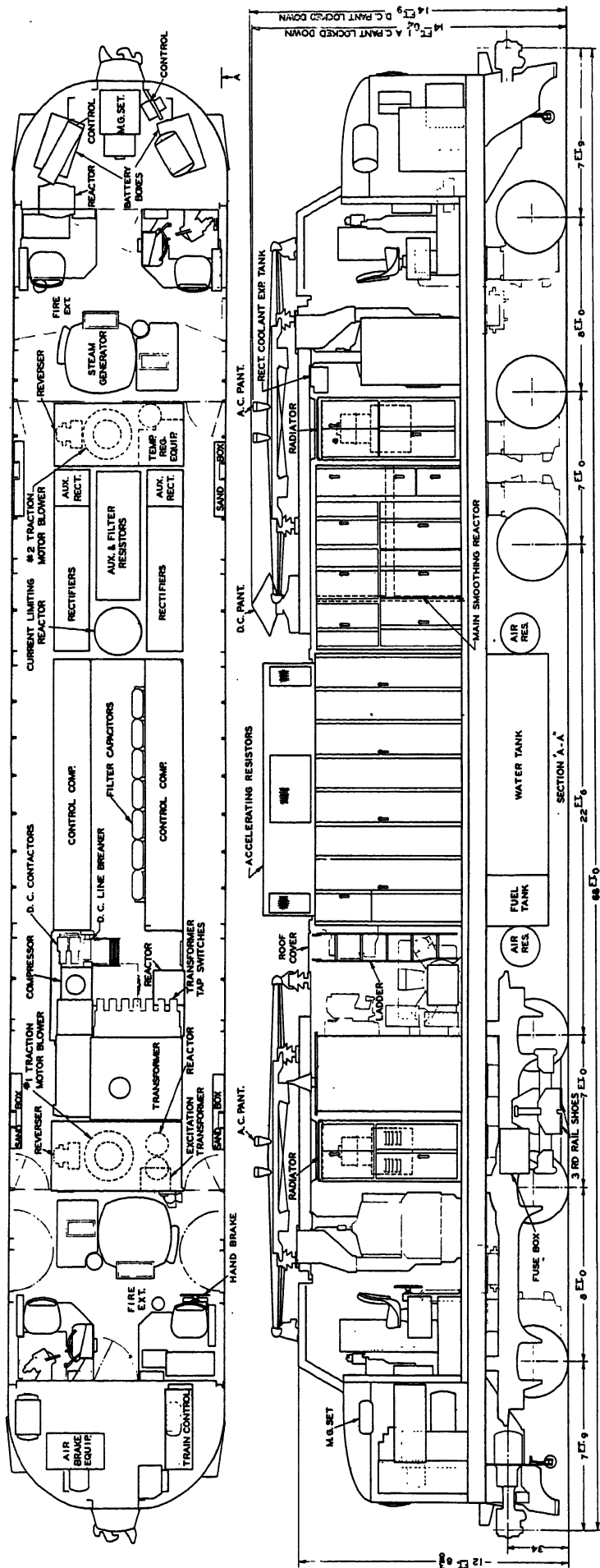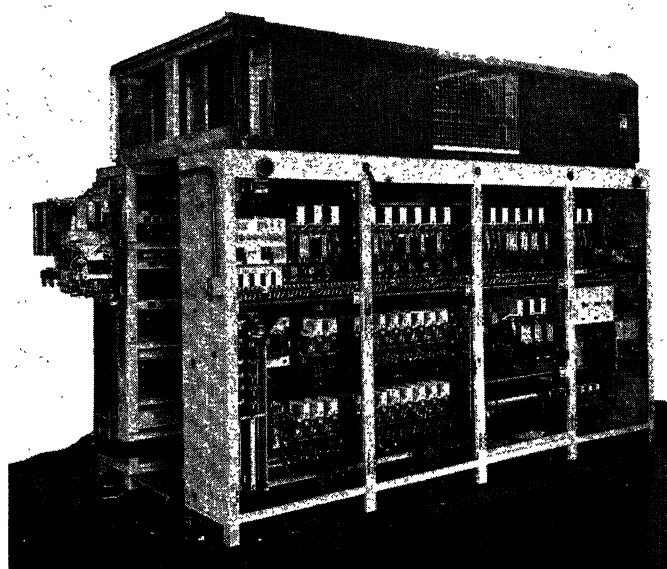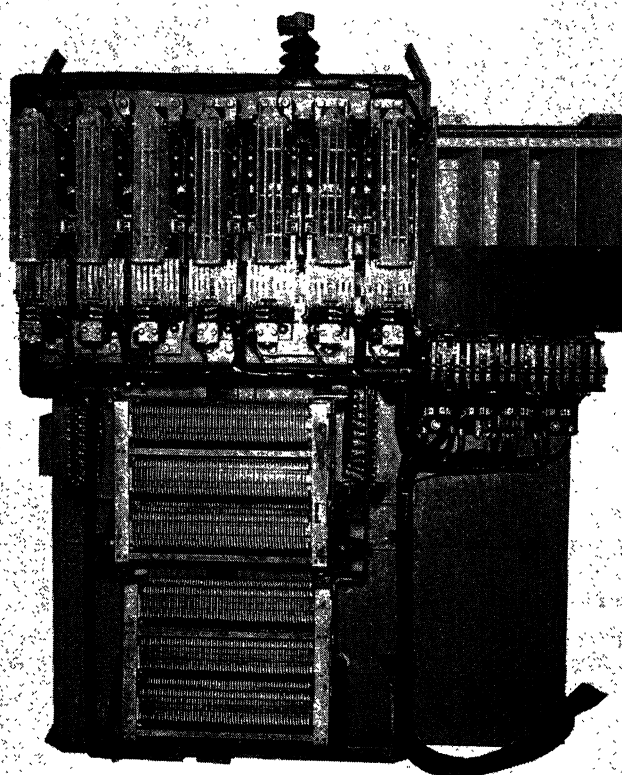
Fig. 4 (left). Main transformer group ready for placement in the locomotive

Fig. 5 (above). Control unit consisting of two control cabinets with accelerating resistors and blower mounted above

lights, marker lights, and number lights are built into the nose sheets. A combined fuel and water tank for train-heating supplies is hung below the cab underframe between the trucks.

Arrangement of equipment on the roof provides for an a-c pantograph mounted on a hatch cover at each end. These hatch covers are built as ducts. Ventilating air for the equipment is taken in from the side of these ducts at the roof level. The center portion of the roof is solid and carries the accelerating resistors and blowers. Both ends of the cab are streamlined, with front windows above the curved hood and windows at the same level on the side to give maximum visibility. Side doors at both ends provide entrance to the operating cabs. Access to the roof is by means of an inside ladder and roof opening.

The cab is carried on two 3-axle, swing-bolster trucks; see Fig. 2. All axles are motored. To provide room for the motor on the center axle, the centering plate is located between axles 1 and 2. Spring-loaded sliding plates on each side of the truck frame between axles 2 and 3 also carry load and provide uniform distribution of weight on the axles.

Inside equalizers, supported on top of the roller-bearing journal boxes, carry the cast steel truck frame on helical springs. The frame in turn carries the swing bolster and center plate on four swing links, spring plank, and elliptical springs. Foundation brake gear is designed for 75-per-cent braking. Clasp brakes are employed with one cylinder for each wheel. Each cylinder is fitted with an automatic slack adjuster.

The weights, dimensions, and ratings of the locomotive are as follows:

Weight

Total locomotive fully loaded, 348,000 pounds
Per driving axle, 58,000 pounds

Dimensions

Track gauge, 56$^1/_2$ inches
Length inside knuckles, 68 feet
Height over cab roof, 12 feet 8$^3/_8$ inches
Height d-c trolley locked down, 14 feet 9 inches
Height a-c trolley locked down, 14 feet 8$^1/_4$ inches
Width over cab sheets, 9 feet 11$^1/_8$ inches
Width over-all, 10 feet 5$^5/_8$ inches
Total wheel base, 52 feet 6 inches
Rigid wheel base, 15 feet 0 inches
Length between center plates, 44 feet 0 inches
Wheel diameter, 40 inches
Coupler height, 34 inches
Clearance motor gear case to rail, 4$^1/_2$ inches
Minimum curve, 288 feet, 20 degrees.

Supplies

Train-heating water, 1,800 gallons
Train-heating fuel, 400 gallons
Sand, 20 cubic feet

Ratings

Tractive effort at 25-per-cent adhesion, 87,000 pounds
Tractive effort continuous, 34,100 pounds
Speed at continuous rating, 44 miles per hour
Power at continuous rating, 4,000 horsepower
Maximum speed, 90 miles per hour
Train heat boiler steam cap, 5000 pounds per hour

Location of Apparatus

Principal pieces of apparatus are shown in Fig. 3. Generally speaking, the small pieces such as batteries, train control, and air-brake equipment are located in the end hoods. Here they are readily accessible and easily connected to the equipment at the engineman's position.

Standard *24RL* air-brake equipment is used. An air compressor with a rating of 200 cubic feet per minute supplies the air. The two storage reservoirs having a combined capacity of 75,000 cubic inches are located beneath the cab underframe, one at each end of the fuel and water tank.

Duplicate operating cabs with raised platforms at the engineman's and fireman's positions are located at each end

of the main apparatus compartment. Controls, brake valves, and other equipment are arranged for the safety and convenience of the crew. A steam generator with a capacity of 2,500 pounds of steam per hour is located in each cab. These two generators can be operated together or independently to meet the train-heating requirements.

A bulkhead with doors on each side separates the operating cab from the main apparatus compartment. An equipment blower at each end of the apparatus compartment furnishes ventilating air for the traction motors, transformer, rectifiers, and reactors. It also ventilates the rectifier cooling radiators. Air enters by means of the ducts in the hatch covers, passes through the rectifier cooling radiators into the equipment blowers which discharge it into a duct in the cab underframe. From here it is distributed to the apparatus as required.

The main transformer, Fig. 4, is designed to utilize the full height of the cab. The high-voltage bushing on top protrudes through the hatch cover to confine the 11,000-volt circuits to the roof. Secondary taps and tap switches are mounted on one end of the tank. The transformer is liquid-cooled by means of Pyranol.* A motor-driven pump circulates the liquid through the tank and windings to a cooler on the right side of tank. Ventilating air from the duct in the cab underframe is forced through the cooler.

All of the d-c control as well as the auxiliary and changeover switches are located in the two control cabinets; see Fig. 5. Each cabinet faces a side aisle

---

* Registered U.S. Patent Office.

from which the equipment can be easily inspected and maintained. A center aisle between the cabinets gives access to the rear of each and to the filter capacitors located there.

The control cabinet structure supports the accelerating resistors and blowers on the roof. This arrangement keeps the length of the leads between the resistors and the switches at a minimum. The resistors are of the ribbon type and are ventilated by means of motor-driven blowers connected across the resistors. With this scheme, the amount of ventilating air is proportional to the voltage across the resistor.

Twelve main circuit rectifiers and four auxiliary rectifiers as well as the firing circuit equipment are arranged in the two rectifier cabinets; see Fig. 6. The rectifier-cooling system pump with its controls is located in the lower part of the blower compartment adjoining the rectifier cabinets. Standard 8-inch rectifier tubes have been modified for this application to provide heavier anode construction and changes in the mercury pool to assure ignition contact under the most adverse conditions of locomotive movement. Cooling water is circulated through the tubes, control valves, and air-cooled radiators by means of a motor-driven pump.

Two reactors, located between the two rectifier cabinets, are air-cooled from the main ventilating duct. One is an air-core type of reactor with six separate coils for limiting the arc-back current in each of the rectifier circuits. The other is an iron-core reactor used to smooth the current ripple in the rectifier output to the traction motors. This is necessary for proper commutation of the motors. The auxiliary motors also have smoothing reactors for the same purpose.

Electric Circuits

The main power circuit arrangement is shown in Fig. 7 and the propulsion circuits are illustrated in Fig. 8. In the a-c zone, power is supplied to the main transformer primary from the overhead trolley at 11,000 volts, 25 cycles. It is collected by means of two pantographs mounted on the roof. Either or both can be used, as the occasion demands. From the transformer secondary, power is supplied through tap switches and current-limiting reactors to 12 main rectifier tubes arranged in three separate bridge-connected circuits. Power output from the rectifier supplies six traction motors through the changeover switches, the main smoothing

---

Fig. 6 (below). Rectifier cabinet—two per locomotive.

Fig. 7 (right). Schematic diagram of main power supply circuits.
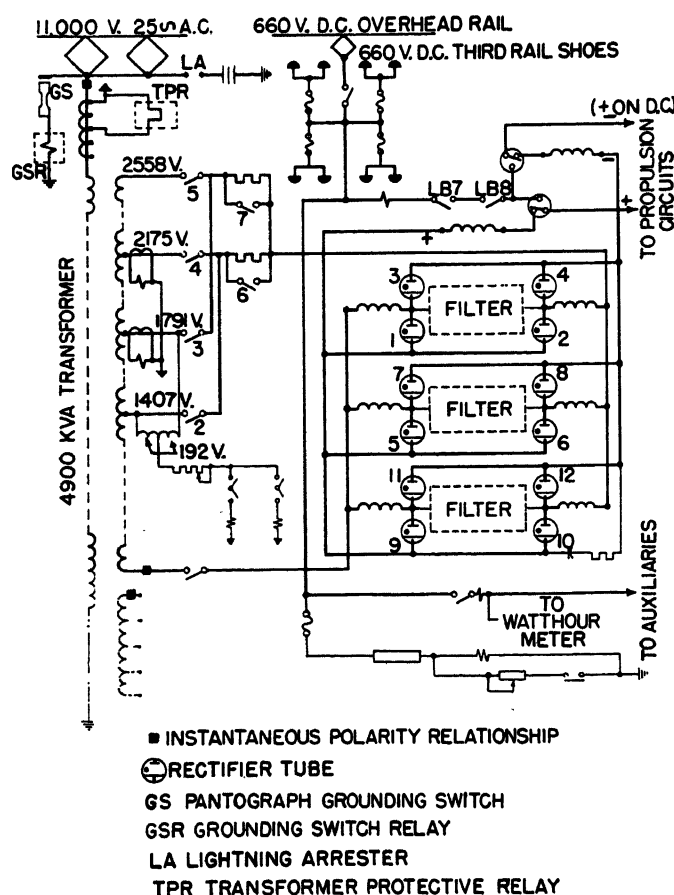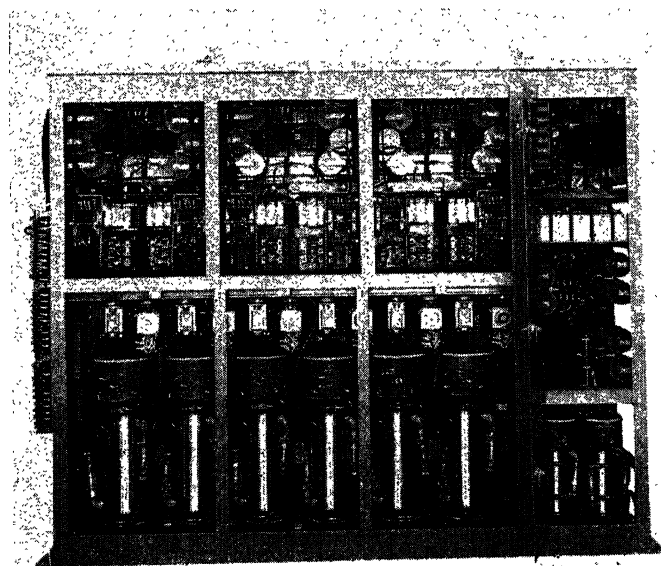
■ INSTANTANEOUS POLARITY RELATIONSHIP
⊖RECTIFIER TUBE
GS PANTOGRAPH GROUNDING SWITCH
GSR GROUNDING SWITCH RELAY
LA LIGHTNING ARRESTER
TPR TRANSFORMER PROTECTIVE RELAY

TO POWER SUPPLY +

TO POWER SUPPLY

TO AUXILIARIES

I(+ ON D.C.)

LOCO. PLATFORM
MOTOR GROUND BRUSHES
RAIL RETURN
WATTHOUR METER

Ⓢ MOTOR CUTOUT SWITCH INTERLOCK | Ⓒ CHANGEOVER SWITCH INTERLOCK | ⤳ MOTOR OVERLOAD RELAY ⊣⊢ REVERSER CONTACT

Sequence table (STEP):

	A. C. NORMAL OPERATION
19	
23	
27	
31	
35	

	D. C. NORMAL OPERATION
19	
31	

Column headings: 1, 2, 3, 4, 5, 6, 7, LB1, LB2, LB3, LB4, LB5, LB6, S1, S2, S3, S4, S5, S6, P1, P2, P3, P4, P5, P6, R1, R2, R3, R11, R12, R13, R21, R22, R23, R24, R31, R32, R33, R34, R41, R42, R43, R44, R50, R51, R52, R53, R54, R61, R62, R63, R64, R65, GC1, GC2, GC3, GCO2, REC, TR1, TR2, TD1
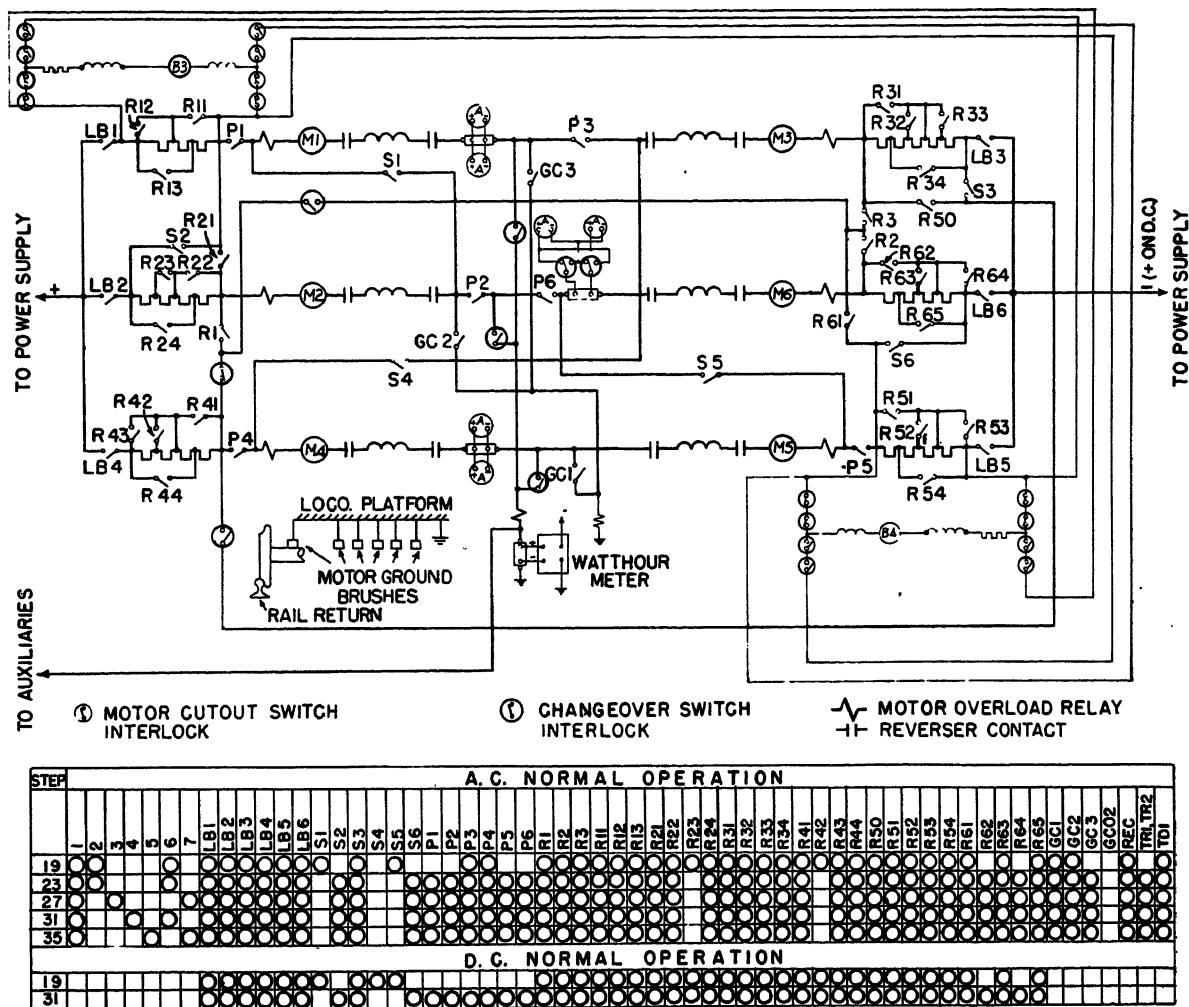
reactor, line switches, and accelerating resistors. Arrangement of the traction motor circuit provides two motor combinations: three in series two in parallel on the 1,407-volt tap, and two in series three in parallel on the three higher voltage taps. Voltage variation on the traction motors is obtained by means of the d-c accelerating resistors in the traction motor circuit in combination with the four transformer taps.

In the d-c zone, power is supplied from the third rail at 660 volts. It is collected by means of third-rail shoes located on each side of both trucks. Connections are made from fuses through two line circuit breakers and a changeover switch to the accelerating resistors and traction motors. Arrangement of the traction motor circuit provides two motor combinations: two in series three in parallel for starting and low-speed operation, and full parallel for high-speed operation. Locomotive characteristics are shown in Fig. 9. Changeover switches located in the main power and auxiliary circuits and controlled from a single switch at the engineman's position permit setting up the circuit for either a-c or d-c operation.

Auxiliary motors are designed to operate on 660 volts d-c. In the a-c zone, an auxiliary rectifier supplies this power from a separate winding on the secondary of the main transformer. Four standard ignitron rectifier tubes arranged in a diametric connection are used. They are water-cooled, as are the main rectifier tubes, from a common

AMPERES PER MOTOR

SPEED — MPH

FS-1 AMPERES

MAXIMUM POWER

CONTINUOUS RATING 4000 HP

TRACTIVE EFFORT - THOUSAND POUNDS

PERCENT ADHESION

SPEED - TRACTIVE EFFORT CURVE
RECTIFIER PASSENGER LOCOMOTIVE
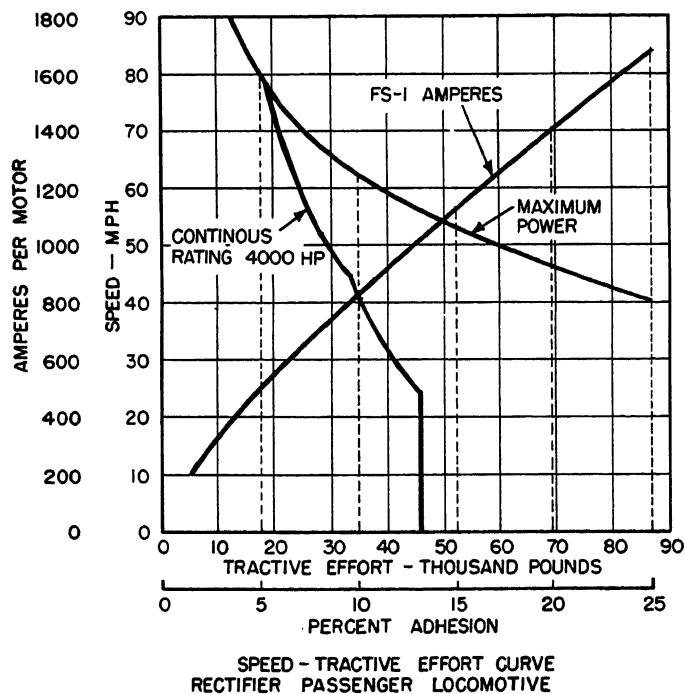
Fig. 9 (left). Locomotive characteristic curves

Table I. Comparison of A-C and Rectifier Locomotives

	2-C+C-2 oo000+00000	C—C 000 000
Year built	1937	1954
Type	a-c commutator motor	rectifier
Service	passenger	passenger
Horsepower at continuous rating	4,000	4,000
Length over knuckles, feet	77	68
Weight		
Trucks	159,000	72,000
Cab and platform	72,000	93,000
Equipment	180,000	156,000
Supplies	21,000	27,000
Total	432,000	348,000
On drivers	270,000	348,000
Per driving axle	45,000	58,000
On guiding axles	162,000	
Per guiding axle	40,500	

cooling system. Smoothing reactors are employed in each auxiliary motor circuit as well as in the main motor circuit.

Battery charging and control power is provided by means of a motor generator set. The generator is designed for 75 volts and supplies the electric equipment of the train-heating boilers, the rectifier cooling-pump motor, battery charging, and control.

Protection for electric equipment is of two general kinds: the overload type for the transformer, traction motors, and auxiliary motors; and the ground relay type for both the main and auxiliary circuits. The rectifiers are protected by means of current-limiting reactors in each circuit and an arc-back relay which removes excitation from the tube firing circuits.

Mechanical Design

These locomotives are the first built, for present New Haven electrical passenger service, with all weight on drivers. Two principal factors have contributed to this favorable design: the conditions of loading on the Park Avenue viaduct, and the type of truck used.

The a-c commutator motor type of locomotives built for this service in 1937 weigh 432,000 pounds. This load is carried on two 3-axle, articulated driving trucks and on two 2-axle guiding trucks, making a total of ten axles. The high load concentration resulting from the close truck and axle spacing limited the axle loads to 45,000 pounds.

In a recent building program the Park Avenue viaduct was modernized and its load-carrying capacity increased. When the rectifier locomotive was proposed with all weight on two widely spaced 3-axle, swing-bolster trucks, it was approved for 58,000 pounds per axle. With six axles, this means that the total locomotive weight was limited to 348,000 pounds. A weight comparison of these two types of locomotives, Table I, shows that the new locomotive weighs 84,000 pounds less than the locomotive built in 1937 and has 74,000 pounds more weight on drivers. The weight reduction

is chiefly in the truck, with the reduced equipment weight offsetting the increase in cab and platform weight.

Summary

Recent application of sealed tube rectifiers on locomotives and cars is of significant importance in railway electrification. In its present state of development, the rectifier tube represents a high-power package that can be used effectively in the transportation field. Its ability to convert commercial frequency a-c power to d-c power for traction opens up many new possibilities for existing as well as future railway electrification.

There is also a program of intensive development and search for new materials and methods. New rectifier materials show promise of even more power in smaller packages with simple control and lower losses. Prospects of reducing costs are also good. These possibilities and the fact that highly developed components such as traction motors and control equipment are available from the Diesel-electric locomotive offer means for reducing the first cost of electric locomotives—an item which has been a barrier to further electrification.

Weight of electric equipment for the rectifier-type locomotive is approximately 10 per cent less than that for the a-c commutator type of locomotive of the same horsepower rating. Railroad modernization combined with progress in both the electrical and mechanical aspects of locomotive design has produced a new locomotive which has 29 per cent more weight on drivers, yet weighs 19 per cent less than the locomotive built for the same service in 1937.

Discussion

Jacob Stair, Jr. (Pennsylvania Railroad, Philadelphia, Pa.): Mr. Gowans' paper is instructive, and it is anticipated that the operation of these new locomotives will be as successful as that of the New Haven rectifier multiple-unit cars.

The advantages of rectifier railway equipment have been discussed at length: the more rugged and simple construction of the d-c motor compared to the single-phase commutator motor; the application of parts now widely used in Diesel-electric locomotives, thereby permitting the manufacturers to build larger quantities of a smaller number of items and enabling the railroads to carry a smaller stock of spare parts; and the ability to change the system from 25 to 60 cycles at a relatively small cost.

As the germanium rectifier is more

efficient than the rectifiers now applied to railway equipment, the use of germanium rectifiers would reduce the amount of cooling now needed to dissipate the heat losses. Furthermore, the germanium rectifier does not require such extensive control circuits. Therefore, it would be helpful to have such rectifiers available which could be applied to railway equipment.

H. F. Brown (Consulting Engineer, New Haven, Conn.): The author omits, in the section entitled "Application," any reference to the two rectifier-type locomotives designed for the Pennsylvania Railroad which were placed in service early in 1952.[1,2]

The New Haven locomotives described in this paper differ materially from those now in service in that the rectifiers are used two in series in a bridge-type circuit, with three bridge circuits in multiple, all on a single secondary winding of the main

transformer at a high voltage. Thus, this New Haven locomotive partakes somewhat of the 3,000-volt d-c design in so far as the switching contactors are concerned, since the potentials switched will at times be close to 2,800 volts. This locomotive could, with but minor changes to the d-c motor connections, operate on an a-c extension of a 1,500-volt d-c system. This is hardly of interest in this country, but might interest some of the European railway people. Whether it was necessary or not to have the circuits of this design for the New Haven operation on 600-volt d-c third rail, it shows the versatility of the rectifier-type locomotive.

The high secondary voltage and its rectified direct voltage would appear to present some operating problems new to the New Haven Railroad, which has not had to deal with voltages higher than 700 or 800 volts heretofore on switched circuits inside the locomotive.

The use of two motors in series, not mechanically coupled, would be considered undesirable by some locomotive designers and users because of the liability of slipping any one driving axle at starting on slippery track or, more important, slipping at high speed when adhesion is low because of a low rail joint or other track imperfection causing one pair of drivers to lose traction

momentarily. It is to be hoped that each axle of this locomotive is adequately protected by a slip relay.

The fact that rectifier-type railroad motive power has now been built by the two leading manufacturers of electric apparatus in this country is of great importance. It indicates the possibility of further application and of further railway electrification.

REFERENCES

1. THE PENNSYLVANIA RAILROAD IGNITRON-RECTIFIER LOCOMOTIVE, C. C. Whittaker, W. M. Hutchison. *AIEE Transactions*, vol. 71, pt. II, Jan. 1952, pp. 37–47.

2. SOME APPLICATION PHASES OF THE IGNITRON RECTIFIER LOCOMOTIVES ON THE PENNSYLVANIA RAILROAD, F. D. Brown. *AIEE Transactions*, vol. 73, pt. II, July 1954, pp. 128–35.

Graphic Aids for Calculating Rectifier Locomotive Performance

R. D. CHARLTON
MEMBER AIEE

THE SUCCESS of any product is often determined by its ability to give customer satisfaction when operated under a wide range of conditions. All too often, these conditions differ appreciably from the specification upon which performance guarantees are based. To assure product satisfaction, the ability to project correctly the knowledge of product behavior from known to unknown regions is essential.

Electric locomotives are usually sold to a specification, and performance is calculated over a given profile on the assumption that voltage at the trolley is the specification value. The accuracy of this assumption depends upon the system itself, and the nature, magnitude, and relative location of other system loads. To be confident that a locomotive, as specified, will give satisfaction when applied, it is necessary to put the locomotive and the system together early in the design stages. In the process of designing some 11,000-volt 25-cycle rectifier locomotives, it became desirable to investigate the locomotive performance, recognizing the effect of system constants and normal operating procedures upon this performance.

Because graphic methods offer a way of organizing knowledge, permitting systematic extrapolation into unknown areas, a generalized graphical method was developed for calculating speed-tractive effort curves. This method was calibrated by comparison with test

values obtained for the same conditions. In addition to determining performance, this method has assisted in visualizing the relative effect on performance of the several factors involved.

General Method of Calculating Performance

To explain the development of the method of calculating rectifier locomotive performance, a succession of locomotives will be considered, ranging from an idealized locomotive to an actual locomotive. In all cases a single motor equivalent is considered.

IDEALIZED LOCOMOTIVE

The circuit for the idealized locomotive is given in Fig. 1; alternating and direct voltages and current wave forms are indicated. In this circuit the assumptions are:

1. Infinite inductance in motor circuit (d-c side of rectifier).
2. No resistance.
3. No supply machine or line reactance (a-c side of rectifier).
4. No rectifier or brush drop.
5. e per rpm is proportional to motor current.

For an applied voltage

$$e = e_m \sin \omega t \qquad (1)$$

Since there are no losses in the system or in the rectification process, the average voltage applied to the motor becomes

$$E = \frac{e_m}{\pi} \int_0^\pi \sin t\, d(t) \qquad (2)$$

which is

$$E = 0.63 e_m \qquad (3)$$

The net voltage around the motor circuit is 0, so

$$E = E_c \qquad (4)$$

where

$$E_c = k_c SI \qquad (5)$$

Combining equations 3, 4, and 5 and solving for speed

$$S = \frac{0.63 e_m}{k_c I} \qquad (6)$$

So the calculation of speed poses no problem.

IDEALIZED CIRCUIT WITH RESISTANCE

Next, consider the addition of resistance terms, tube, and brush drops. The addition of resistance terms does not materially alter the procedure just outlined. The sum of voltages around the circuit is still zero. The alternating current has the same average value as the motor current. One ohm in the a-c side will cause the same average voltage loss as 1 ohm in the d-c side. It is convenient to lump resistance terms

$$E = IR + E_c \qquad (7)$$

Nomenclature

e = instantaneous voltage
e_m = peak alternating voltage
e_r = rms alternating voltage
E = average voltage
E_c = motor counter electromotive force
E_D = tube and brush drop voltage
E_X = commutation loss voltage
I = average motor current
i = instantaneous a-c amperes
R = resistance, ohms
X = reactance, ohms

f = frequency, cycles per second
ωt = time, degrees
$\omega = 360f$
S = motor revolutions per minute (rpm)
θ = commutating period, degrees
f_a = voltage ratio, $f_a = E/e_r$
f_r = voltage ratio, $f_r = e_r/e_m$
k_c = motor counter electromotive force constant (from motor characteristic), $k_c = E_c/SI$
a = reactance ratio
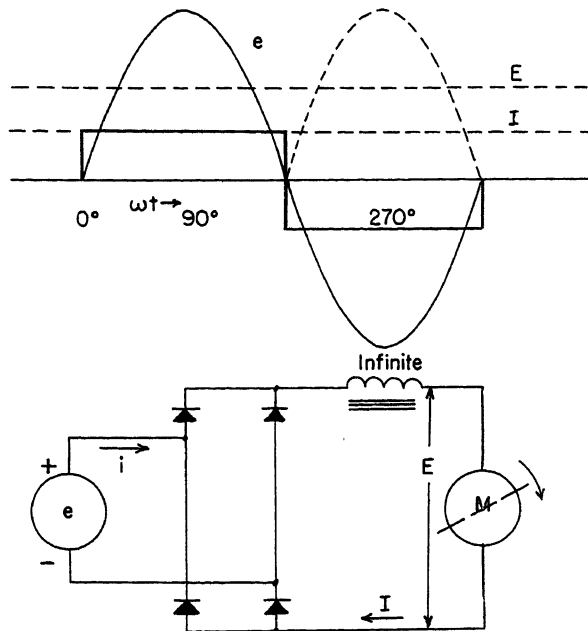
Fig. 1 (left). Idealized locomotive circuit
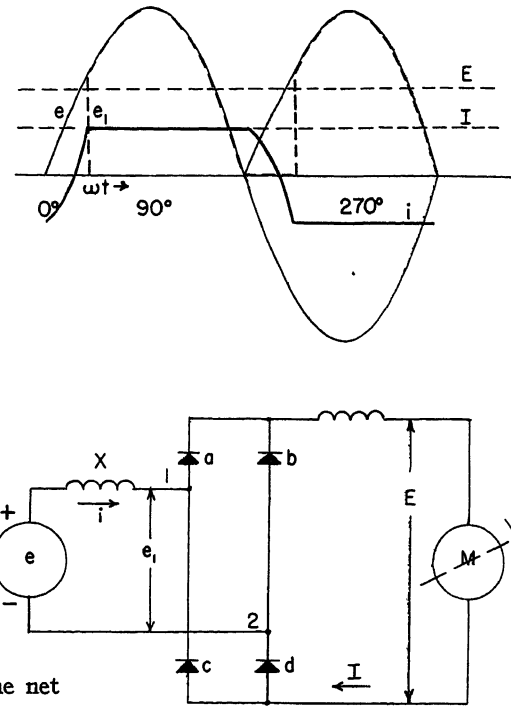
Fig. 2 (right). Simplified commutating circuit

Again solving for speed

$$S=\frac{0.63e_m-IR}{k_cI} \tag{8}$$

The term k_cI represents counter electromotive force per rpm at a value of current I. In actual practice the value of this term would be obtained from a motor counter electromotive force per rpm characteristic for the value of current in question.

Tube and brush drops are of the same class as the current-resistance drops and may be included in equation 8 as follows

$$s=\frac{0.63e_m-IR-E_D}{k_cI} \tag{9}$$

Equation 9 represents speed for rectifier locomotive operation on infinite bus, for the conditions assumed.

LOCOMOTIVE ON INFINITE BUS

To obtain a circuit which has some direct application, consider the addition of reactance in the a-c circuit.

Refer to Fig. 2 for circuit under consideration, and voltage and current wave forms. The inclusion of a-c reactance X brings a new factor, commutation, into the picture.

Consider the voltage and current wave forms at $\omega t = 180$ degrees. At this instant, rectifiers a and d are conducting, and current throughout the system is I. The voltage rise was from 2 to 1, it is going to be from 1 to 2. As point 2 becomes positive with respect to 1, commutation begins; tubes c and b conduct. From this time until tubes a and d cease conducting, the only voltage which can exist between point 1 and 2 is drop across the conducting rectifier tubes. The net drop across the tubes is so small compared to the supply voltage as to be negligible.

During the commutating period when all tubes are conducting, all the supply voltage must appear across the a-c reactance. Also during this period, the motor current remains constant at I amperes, while the line current changes from I and $-I$. To evaluate the relation among the several variables, separate the circuit of Fig. 2 into two parts, as shown in Fig. 3.

By superposition, the current in tubes a and d, i_a and i_d, is the sum of the separate currents

$$i_a=\frac{i}{2}+\frac{I}{2}=i_d \tag{10}$$

where the positive sense is determined by the conducting direction of the tube.

Also

$$i_b=i_c=\frac{I}{2}-\frac{i}{2} \tag{11}$$

where i is the transient current which flows while all tubes conduct. Equations 10 and 11 apply only for positive values (conduction).

The solution for i may be obtained by considering the simple circuit of Fig. 4. The driving voltage is

$$e=e_m \sin \omega t \tag{1}$$

The switch is closed at $\theta=180$ degree, when all tubes conduct at the start of the second half-cycle.

Also

$$e=L\frac{di}{dt} \tag{12}$$

so

$$e_m \sin \omega t=L\frac{di}{dt} \tag{13}$$

or

$$i=\frac{e_m}{X}\int_\pi^\theta \sin \omega t\, d(\omega t) \tag{14}$$

Solving, and substituting the boundary condition that $i=+I$ when $\theta=\pi$

$$i=I-\frac{e_m}{X}(1+\cos \theta) \tag{15}$$

As mentioned earlier, equation 15 holds only for positive values of i_a, i_b, i_c. and i_d, from equation 17, when $\theta = $ arc cos $-(1-(2IX)/e_m)$. To find the time at which this switching of current between pairs of tubes is completed, substitute equation 15 into equation 10; in the limit, $i_a=0$ so

$$i_a=0=I-\frac{e_m}{2X}(1+\cos \theta) \tag{16}$$

Since θ is a third quadrant angle, the minus sign can be dropped and θ is expressed as degrees after voltage 0 for convenience

$$\cos \theta=1-\frac{2IX}{e_m} \tag{17}$$

Physically, energy stored in the a-c reactance at the initial condition of I amperes is returned to the system and then again stored in the a-c reactance, now at $-I$ amperes. All during this period the propulsion equipment has been coasting along on the energy stored by the infinite inductance assumed in that
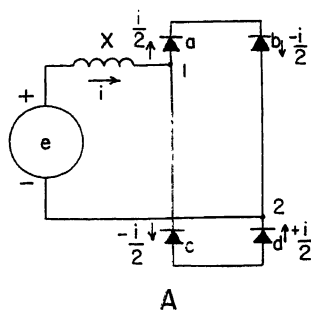
Fig. 3 (left). Commutating currents

A—Alternating
B—Direct
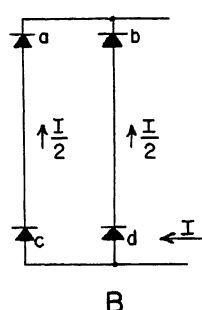
Fig. 5. (right). Locomotive on finite bus

$$f_a = \frac{1+a+\cos\theta(1-a)}{\sqrt{\pi\left(\frac{\pi}{2}+(1-a^2)(1/4\sin 2\theta - \frac{\pi}{90}\theta)\right)^{1/2}}} \quad (27)$$

$$f_r = \left[\frac{1}{\pi}\left(\frac{\pi}{2}+(1-a^2)(1/4\sin 2\theta - \frac{\pi\theta}{90})\right)\right]^{1/2} \quad (28)$$

section of the circuit. The volt-seconds input to the system during tube commutation was absorbed by a-c reactance, and, not being available at the rectifier output, represents a voltage loss or drop as far as performance is concerned.

The volt-second loss expressed as an average is

$$E_x = \frac{e_m}{\pi}\int_0^\theta \sin \omega t d(\omega t) \quad (18)$$

where

$$\theta = \arccos\left(1 - \frac{2IX}{e_m}\right)$$

Integrating equation 18 and substituting limits for t

$$E_x = \frac{2IX}{\pi} \quad (19)$$

This term has the characteristics of resistance. As such it may be introduced directly into equation 9

$$s = \frac{0.63e_m - I\left(R + \frac{2X}{\pi}\right) - E_D}{k_cI} \quad (20)$$

Equation 20 is satisfactory for the calculation of speed where supply is infinite and maximum voltage is known. In general term, equation 20 becomes

$$S = \frac{f_a e_r - E\left(R + \frac{2X}{\pi}\right) - E_D}{k_cI} \quad (21)$$

LOCOMOTIVE ON LESS THAN INFINITE BUS

In actual practice, the circuit looks as in Fig. 5. Normally, voltage e_1, representing a feeder or bus, is maintained at some

Fig. 4. A-c transient loop

predetermined value. For the purpose here, it is assumed that the regulating equipment will maintain an rms voltage, regardless of wave shape distortion caused by the rectifier commutation.

The starting point for calculating speed in the circuit of Fig. 5 is e_1, a known rms voltage. The various average voltage losses in the circuit from the point where e_1 is measured through the traction motors must be subtracted from the average value of e_1 to obtain the average voltage available for speed. In equation 21, the factor relating rms and average voltage must be changed, since what is known is the rms value of a nonsinusoidal voltage, e_1. All other voltage terms in equation 21 can be directly applied; the general reactance X becomes X_B. During commutation, since the same rate of change of current exists throughout the a-c loop, the applied voltage is distributed between X_A and X_B in proportion to their individual magnitudes; see Fig. 6.

If the basic wave is

$$e = e_m \sin \omega t \quad (1)$$

then the voltage wave e_1 can be expressed in two parts:

For values of ωt between 0 and θ

$$e_1 = \frac{X_B}{X_A + X_B} e_m \sin \omega t \quad (22)$$

For values from θ to π

$$e_1 = e_m \sin \omega t \quad (23)$$

Evaluating average and rms values for e_1

$$E = \frac{e_m}{\pi}\left(\int_\theta^\theta a \sin \omega t d(\omega t) + \int_\theta^\pi \sin \omega t d(\omega t)\right) \quad (24)$$

where

$$a = \frac{X_B}{X_A + X_B} \quad (25)$$

$$e_r = \frac{e_m}{\sqrt{\pi}}\left(\int_0^\theta a^2 \sin^2 \omega t d(\omega t) + \int_\theta^\pi \sin^2 \omega t d(\omega t)\right)^{1/2} \quad (26)$$

Integrating equations 24 and 26, solutions for the form factors are

Equations 27 and 28 are plotted for several values of a in Figs. 7 and 8.

To obtain a solution to the general expression for speed, equation 21 where the bus is not infinite becomes

$$S = \frac{f_a e_r - I\left(R + \frac{2X_B}{\pi}\right) - E_D}{k_cI_c} \quad (29)$$

where commutating angle θ is a new parameter and f_a is expressed in terms of θ in equation 27.

I is expressed in terms of θ using equation 17

$$I = \frac{(1-\cos\theta)e_m}{2X} \quad (30)$$

where

$$X = X_A + X_B \text{ and } e_m = e_r f_r \quad (31)$$

I can be expressed as

$$I = \frac{(1-\cos\theta)e_r}{2Xf_r} \quad (32)$$

Now all factors in equation 29 have been expressed in terms of system constants or commutating angle.

To obtain a speed curve the following steps are necessary:

1. Assume a commutating angle θ.

2. Extract form factors f_a, f_r, and from Figs. 7 and 8 for the commutating angle assumed and for correct system reactance ratio a.
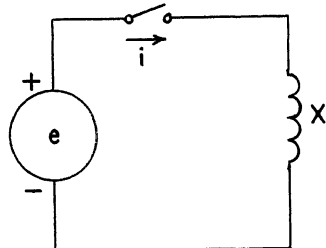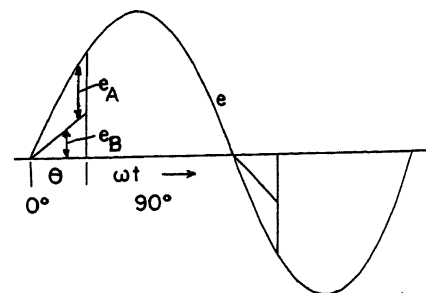
Fig. 6. Trolley voltage on finite bus
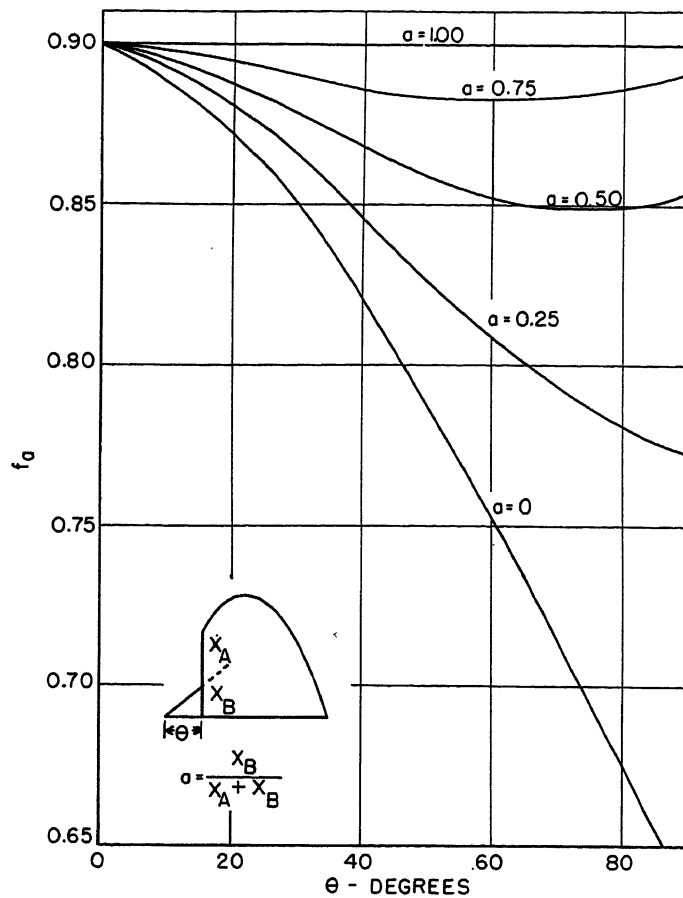
Fig. 7. Ratio of average to rms voltage

Fig. 8. Ratio of average to maximum voltage

3. Compute motor current I from equation 32 using assumed angle and known values of other constants.

4. Compute speed using equation 29 for current calculated in step 3 and known values of other constants.

5. Repeat for different angles until an adequate performance curve is obtained.

Generalized Graphical Method

The process previously outlined, while satisfactory, is slow and requires considerable repetitive calculation. When the entire calculation is looked at with a particular set of system constants defined as the 100-per-cent or per-unit values, the possibility of a graphical development can be considered.

Consider equation 29

$$S = \frac{f_a e_r - I\left(R + \frac{2X_B}{\pi} - E_D\right)}{k_c I} \quad (29)$$

where $k_c I$ represents motor counterelectromotive force per rpm for the current I.

If 100-per-cent values are assigned R, the magnitude of the current-resistance term can be plotted for different per-cent values of R as a family of curves; see Fig.

9(A). A similar set of curves can be drawn $2/\pi(X_B)$, $X_B = a\ (X_A + X_B)$; see Fig. 9(B). The term $a\ (X_A+X_B)$ represents the reactance on the locomotive side of the point at which voltage is maintained. The 100-per-cent value is most conveniently assigned to (X_A+X_B) and the parameter for the $(2X_B)/\pi$ term is $a(X_A+X_B)$ in per cent.

If desirable, resistance and commutation drops can be combined, and the tube and brush drops added. The effect of temperature can be included in the resistance plot. The relation between motor current and f_a can be displayed graphically in a similar fashion, using commutating angle as the common parameter.

Rewriting equation 32

$$\frac{I(X_A+X_B)}{e_r} = \frac{(1-\cos\theta)}{2f_r} \quad (33)$$

Since f_r is expressed in terms of θ and a in equation 28, the right-hand term of equation 33 can be plotted against θ with a as the variable for a family of curves. For any value of θ on this plot, there is a form factor f_a depending upon the value of a. This has already been plotted in Fig. 7. For convenience, the family of curves for $(1-\cos\theta)/(2f_r)$ and f_a may be
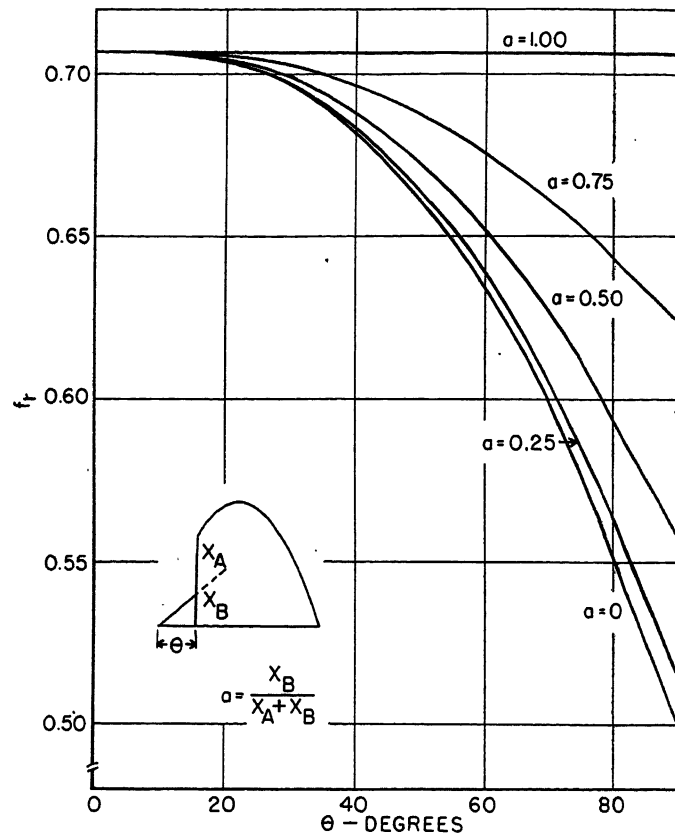
plotted together, as shown in Fig. 11.

The term $1(X_A + X_B)/e_r$ can be evaluated for a range of currents and the 100-per-cent values of (X_A+X_B) and e_r. These quantities are the ordinates which correspond to $(1-\cos\theta)/(2f_r)$ by equation 33. The ordinate scale for the resistance drop curves can be adjusted to bring all current values into correspondence with mating $I[(X_A+X_B)/e_r]$ term. In
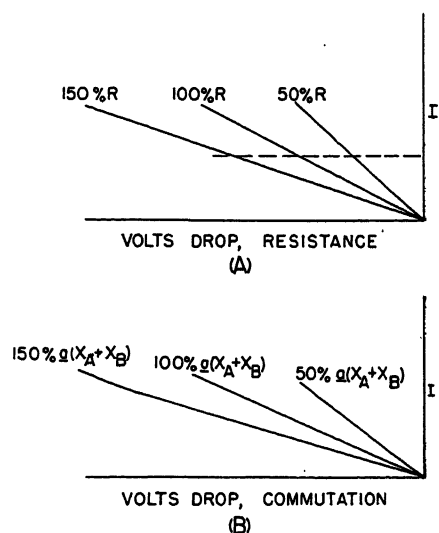
Fig. 9. Voltage loss

A—In resistance

B—During commutation
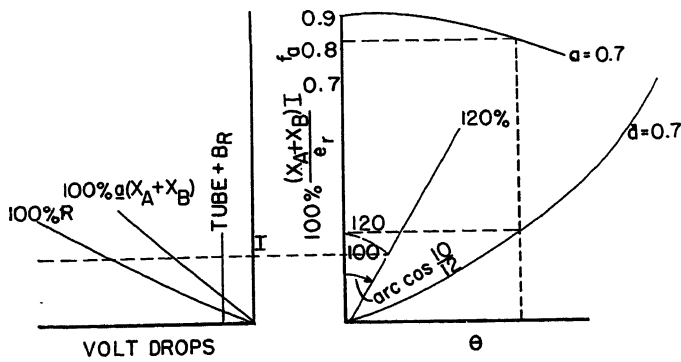
Fig. 10 (above). Example of working graph

this way (by reading across on a horizontal line) resistance drop, commutating reactance drop, and commutating angle for one value of motor current can be read directly. The form factor f_a can be obtained by reading vertically upward from the commutating angle intercept to the correct a value f_a curve.

To use this basic curve on different percent values of (X'_A+X_B) and e_r, it is only necessary to make suitable geometric construction by means of which the assumed value of motor current and the new per cent values of (X_A+X_B) and e_r can be referred to the basic $I[(X_A+X_B)/e_r]$ scale. For instance, if $(X_A+X_B) = 110$ per cent and $e_r = 90$ per cent, the term $I[(X_A+X_B)/e_r]$ will be 120 per cent of the reference for determining the commutating angle. This is indicated in Fig. 10.

With the addition of a means to determine commutating angle for values of e_r and (X_A+X_B) other than base values,

Fig. 12 (right). Comparison of test and calculated results

the over-all chart is essentially complete. The form factor f_a can be determined as before by reading up from the corrected current intercept on the commutating angle curve family. The motor e per rpm characteristic is added as a matter of convenience. A per-cent scale is already available for e_s and E voltages, as form factors are expressed on a per-cent basis. Lines like those employed to step up or step down X and e_r from their base values can be used to determine per-cent E for any value of e_r other than the 100-per-cent value.

Comparison of Calculated Results With Test

To verify the method and provide calibration, a single motor equivalent of a locomotive was set up and tested using full size components. The complete working chart to accompany this example is shown on Fig. 11.

All circuit values have been referred to transformer secondary open circuit voltage and are for a frequency of 25 cycles per second. X_A is the sum of all reactances between the locomotive and the
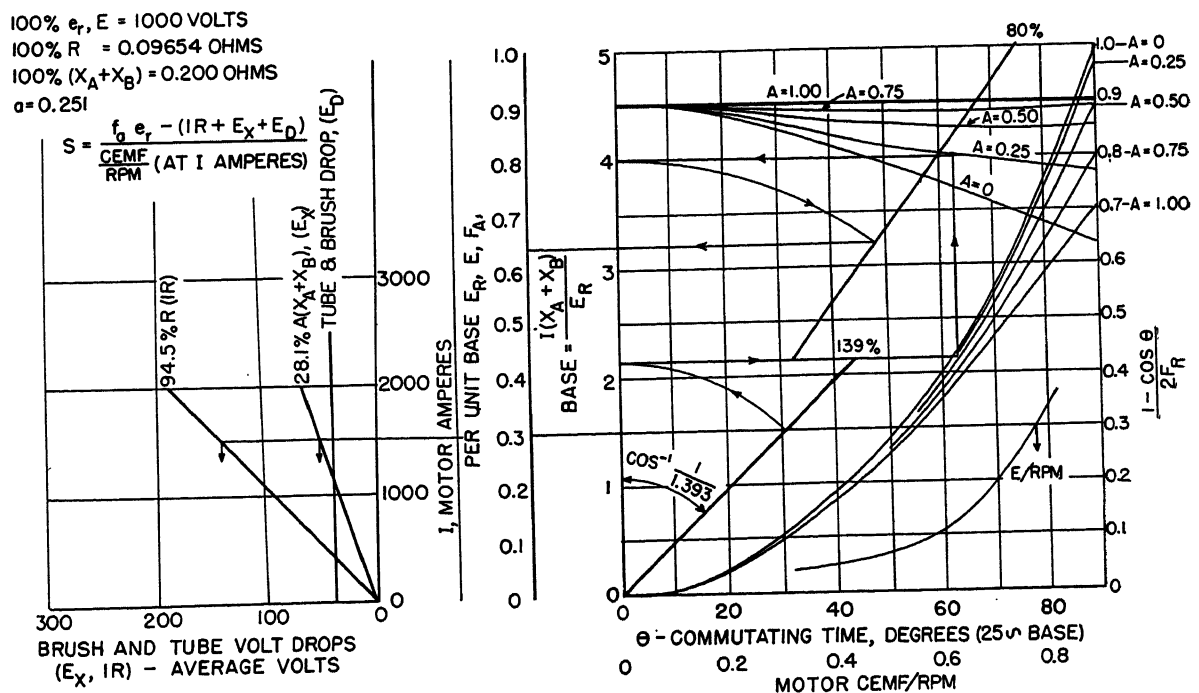
Fig. 11. Complete graphical construction

Table I. Circuit Values

Variable	Test, Ohms	Base	Per-Unit Test
$X_A + X_B$	0.2229	0.200	1.115
a	0.251		
$a(X_A + X_B)$	0.0561	0.200	0.281
e_r	800	1000	0.8
$X_A + X_B$	0.2229	0.200	1.393
e_r (volts)	800	1,000	
R (100 degrees centigrade)	0.09124	0.09654	0.945

supply. Supply machine reactance is taken as 100 per cent of the direct axis subtransient reactance. X_B is the sum of all locomotive reactances not including reactances in the d-c section of the circuit. R is the sum of all locomotive resistances.

The setup under consideration represents a low-voltage, high line-reactance condition. This is abnormal and was selected to determine the error of the calculating method described at its application limit. Circuit values are shown in Table I. Construction is included to show the procedure for evaluating speed at $I = 1,500$ amperes.

The test results for this setup are plotted in Fig. 12. The ratio of calculated to test rpm is also plotted in Fig. 12. The maximum error is seen to be —3 per cent for this extreme condition.

To increase accuracy complicates the procedure. Correction of the resistance line for temperature will improve the accuracy but represents a refinement not in line with the objective originally established. As it is presented, the method will permit the rapid calculation of speed-current curves. For the range of system values likely to be encountered, the calculated speed will be correct to within 3 per cent.

Reference

1. ANALYSIS OF RECTIFIER CIRCUITS, E. F. Christensen, C. H. Willis, C. C. Herskind. AIEE *Transactions*, vol. 63, 1944, pp. 1048–58.

◆

Discussion

L. J. Hibbard (Westinghouse Electric Corporation, East Pittsburgh, Pa.): Experience indicates that, if accurate values of the impedance in the trolley and transmission system including generator impedance can be obtained, the most accurate speed current characteristics can be calculated by using the internal generated voltage at the power supply points. The calculated and actual test values of an existing rectifier motive power application are shown in Table II.

R. D. Charlton: Regarding Mr. Hibbard's discussion, our experience agrees with his.

Table II. Calculated and Actual Test Values

Traction Amperes per Car Test	Test, Miles per Hour	Calculated, Miles per Hour
785	29.8	30.2
632.5	35.7	36.2
530	41.7	41.7
457.5	47.6	48.0
400	54.8	55.0
380	56.6	57.0
358	59.5	60.0
310	71.5	71.0

However, system voltage maintained is not the internal generated voltage of the supply generator(s). System voltage is maintained through voltage regulator action. The regulator and system voltmeter are connected at some point in the system like the trolley-feeder bus in the generating station.

The method presented assumed such a connection and assumed that the system voltage regulator maintained a constant rms voltage at that point as read by the meter. In such a case, reactance on the generator side of the point where voltage is maintained will take a notch out of the voltage wave during rectifier commutation. The average value must be determined for this nonsinusoidal wave with a known rms value. Determination of the relationship between rms and average value of the notched-out voltage wave, under application conditions, helps assure satisfactory locomotive behavior when connected to the customer's system.

An Electric Drive for Rotary Snowplows

A. H. HOFFER
ASSOCIATE MEMBER AIEE

R. E. WILLHITE
ASSOCIATE MEMBER AIEE

AS COMPLETE dieselization of the United States railroads approaches, the problem of replacing steam-powered rotary snowplows increases. Numerous designs have been tried in the past decade but thus far none has satisfactorily met all of the broad requirements. The electric drive here described features high reliability, over-all simplicity, standard interchangeable components, inherent stabilization, and adaptability.

Power Requirements

After evaluating the power required to remove snow from railway tracks at a rate sufficiently high to keep them open, it is little wonder that railway operating personnel felt that they were "sending a boy to do a man's job" when dispatching a steam rotary to clear a line.[1]

The power required to drive a snowplow rotor depends upon: 1. the weight of the snow to be removed per unit of time; 2. the net energy imparted to the snow, and 3. the over-all electrical and mechanical efficiencies of the plow. Hence, the engine horsepower (hp) required by a plow of the centrifugal type may be expressed as

$$HP = \frac{H \times B \times D \times S \times 5,280}{60 \times 33,000 \times e_d \times e_r} \left(h + \frac{v^2}{2g} \right) \quad (1)$$

or

$$HP = \frac{H \times B \times D \times S}{375 \times e_d \times e_r} \left(h + \frac{d^2 N^2}{23,500} \right) \quad (2)$$

where

H = height of the snow to be removed (may vary from 0 to 16 feet)
B = breadth of the cut (may vary from 11 to 15 feet)

D = density of the snow (may vary from 3 to 50 pounds per cubic foot)
S = forward speed of plow, or rate at which the cut is advanced (may vary from 1 to 30 miles per hour)
e_d = over-all efficiency of the drive (may vary from 0.85 to 0.90)
e_r = over-all efficiency of the rotor and discharge funnels (may vary from 0.40 to 0.90)
h = average height that the snow must be lifted in the plow before discharge (may vary from 6 to 12 feet)
V = rotor tip velocity (may vary from 44 to 110 feet per second)
g = acceleration of gravity (32.2 feet per second^2)
d = rotor diameter (may be from 7 to 12 feet)
N = rotor speed in revolutions per minute (may vary from 0.5 to 1.25 times its rated speed)

Since the centrifugal type of snowplow is analogous to a centrifugal fluid pump, it is not surprising that the energy equa-
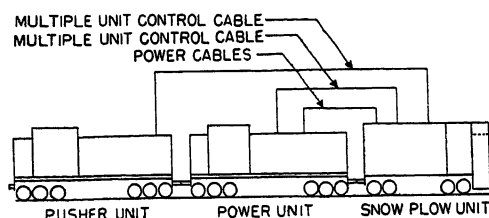
Fig. 1 (left). Head end of snowplow train showing cables between units
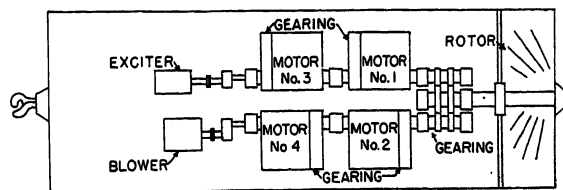
Fig. 3. Schematic arrangement of equipment in snowplow unit

tion of the former has the same form as that of the latter. (See equation 1.) The energy required to drive this snowplow depends upon the changes in potential and kinetic energy imparted to the snow against the restraints of friction, turbulence, and shear. Great care should be taken in the design of the rotor blades and discharge funnels to prevent dissipation of large quantities of energy in useless work.

If α is the angle of discharge of the snow from the rotor chutes, then the range of the rotor (the distance which the snow is cast) is

$$R = (2v^2 \sin \alpha \cos \alpha)/g \qquad (3)$$

When the angle of discharge is 45 degrees the range is a maximum and equation 3 may be written

$$R = v^2/g = 85\left(\frac{d^2}{10}\right)\left(\frac{N^2}{100}\right) \text{ feet} \qquad (4)$$

Equations 2 and 4 shows that both the power and the range of the rotor vary as the square of its speed. Therefore, the power required by the rotor varies directly with the range. Hence, casting the snow further than necessary is a needless waste of power.

From equation 2 it can also be seen that the power required by the rotor varies as the rate at which the cut is advanced. The available power must be divided between advancing the cut and casting the snow. In practice the relationship between these two varies greatly, so the ideal drive should permit full power plant utilization over a wide range of rotor speeds.

The power required by a snowplow also depends upon the length of the line to be cleared in a given time and the rate at which the line is being covered with snow. A rate of snowfall averaging 1 inch per hour is considered quite heavy, but rates up to 4 inches per hour have been recorded in mountain areas.[2] Rate of fall, however, is not always the prime determinant. Depending on wind conditions and nature of the terrain, drifting of even a light snow can have more serious consequences in blocking a line than a high rate of fall.

Finally, the power required by a snowplow depends directly upon the snow density. This varies widely, with its

structure, its previous history, the prevailing temperatures and wind conditions, and the time of settling. The structure may vary all the way from a light crystalline fluff to a heavy, lumpy, wind-packed water-soaked encrustation. The rotor must be able to shear and discharge all types equally well. Fresh snow weighs between 3 and 7 pounds per cubic foot; settled snow can easily weigh as much as 30 pounds per cubic foot; and wet snow, packed by slides and avalanches, has been known to weigh as much as 50 pounds per cubic foot.[2]

System Requirements

Besides possessing sufficient power, a good electric snowplow must fulfill the following generally accepted requirements:

1. It should be sturdy, simple and reliable, both electrically and mechanically.

2. It should be readily adaptable to any standard Diesel-electric road locomotive unit.

3. Its use should require no alterations of the equipment in the power unit.

4. It should be made up of standard interchangeable locomotive components.

5. It should be inherently self-protective against overloads and overspeeds.

6. It should be able to operate gently and smoothly, with no sudden step functions or discontinuities, over its entire speed range.

7. It should be able to adjust itself to meet sudden fluctuations in load.

8. It should be able to load the engine of the power unit to its rated capacity over a wide range of rotor speeds.

9. It should operate equally well in either direction of rotation.
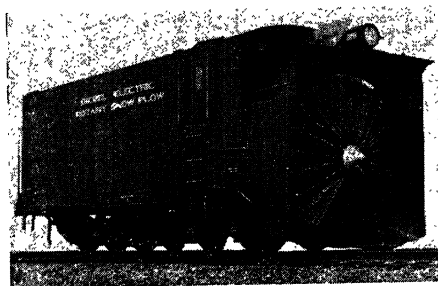
Fig. 2. Conventional rotary snowplow fitted with electric drive

10. The operator should be able to control both the power and the pusher units remotely (multiple-unit operation).

11. The plow unit should be able to withstand the rotor reaction torque and have good tracking ability.

General Description of Electric Drive and Components

A thorough study led to a simple design which satisfactorily meets all of the generally accepted requirements. A stabilizing exciter causes the plow-drive motors to deliver the desired speed-torque characteristics at all times. The manner in which this exciter operates is the novel feature of the drive: by automatically regulating the strength of the drive motor fields it controls the speed of the rotor for any load. The operator has only to see that the plow is moving forward at a speed which results in the snow being cast or farther than is required.

The head end of the snow plow train, including the plow, the power unit, the pusher unit, and the electric connections between them, is sketched in Fig. 1. A conventional centrifugal type of snowplow is shown in Fig. 2. Fig. 3 is a suggested layout, showing rotor, gearing, drive motors, stabilizing exciter, and traction motor blower. The basic wiring diagram of the drive is shown in Fig. 4. Motors 1, 2, 3, and 4 are standard series traction motors with standard gearing. Although four motors are shown, the basic system will work equally well with fewer larger motors capable of handling the rated hp. The stabilizing exciter E_s is an auxiliary d-c generator, rated 25 kw at $37\frac{1}{2}$ volts.

The output of this exciter depends upon its speed and its net excitation. The speed is directly proportional to the speed of the rotor, making the exciter immediately speed sensitive. Its net excitation is the resultant of the shunt and differential field excitations. The shunt field excitation is constant (fixed by the setting $R4$) and determines the no-load speed of the rotor. The differential field excitation (which varies directly with load current through $R1$) determines the speed of the rotor when under load. Simply by

Hoffer, Willhite—Electric Drive for Rotary Snowplows
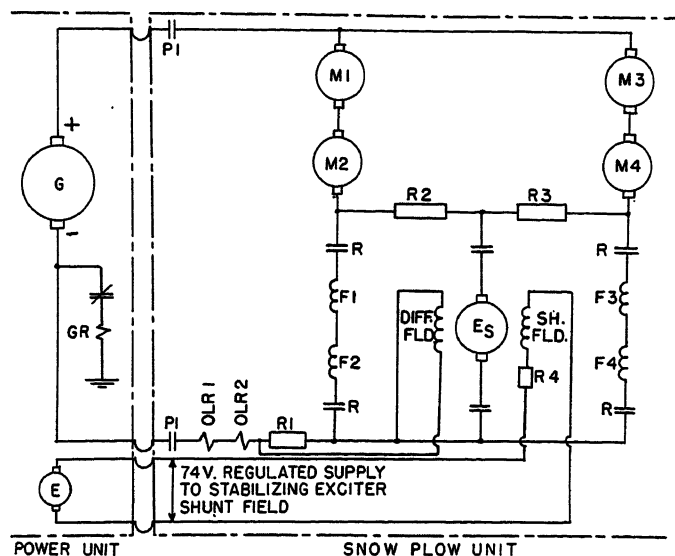
Fig. 4 (above). Schematic diagram of circuits for electric drive

Fig. 5 (right). Output curves for plow powered by a 1,600-hp power unit

adjusting R1 and R4, the speed-torque characteristic of the plow drive can be made to match the volt-ampere characteristic of any standard power unit. Once these adjustments have been made the operation of the drive becomes completely automatic.

R4 is a small resistor panel made up of standard wire wound resistor tubes. R1 consists of two standard field shunting resistors connected in parallel. R2 and R3, the resistors which limit the exciter current during starting, are each a half of a locomotive-type edgewise wound resistor. A standard reverser R allows the drive to operate in either direction of rotation. The use of a small number of standard locomotive components, simply connected, tends to make this drive sturdy and reliable.

Operation

The plow may be equipped to provide complete control of the speed and direction of the snowplow train from the plow cab by means of multiple unit connections, or these functions may remain in the pusher unit as in the past. The power input to the rotor and its direction of rotation are regulated and controlled by a controller similar to that found on locomotive B units.

After the motor circuits of the power unit have been separated from the generator, the extension power leads to the snowplow motors are connected in their places. Additional auxiliary power for the plow is furnished by connecting extension leads to the auxiliary power system of the power unit.

Once the snowplow train has been made up with the necessary power,

auxiliary and control connections, it may be moved to the site of operation. Track conditions ahead can best be observed when operating from the snowplow cab and the forward speed of the plow varied to maintain full load while casting snow the required distance.

To bring the rotor up to maximum speed the snowplow operator notches his controller out all the way; see Fig. 5. As the train moves forward into the snow, the rotor speed drops slightly until the plow has loaded the power unit engine to its rated output; see Fig. 6. Any further demand for increased rotor torque results in a reduction of rotor speed. Although this reduces the distance the snow is cast, the output of the power unit engine remains at a maximum. Therefore, as shown in Fig. 5, the distance the snow is cast can be controlled by varying the forward speed of the train. Hence, the snowplow drives utilizes the maximum output of the power unit engine over an extremely wide range of rotor speeds.

All other conditions being constant, the rotor speed decreases as the forward speed increases. As the rotor speed approaches one half of its rated value, the rotor has difficulty in clearing itself and is in danger of stalling. (Stalling occurs when the snow no longer possesses sufficient kinetic

energy to clear the chutes.) Fig. 7 shows that the line current increases as the rotor speed decreases. Therefore stalling is prevented by setting OLR1 to trip at a predetermined value of generator current.
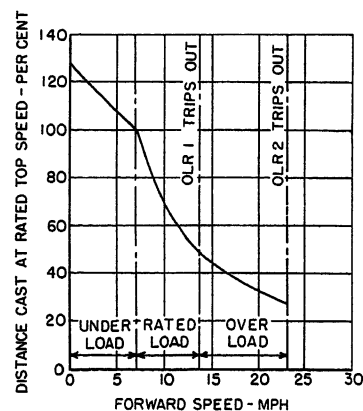
Fig. 6. Speed-distance cast curve for plow when power input is 1,600 hp

Height of snow, 3 feet
Breadth of cut, 12 feet
Density of snow, 10 pounds
   per cubic foot
Drive efficiency, 85%
Rotor efficiency, 68%
Average lift in plow, 6.8 feet
100% distance cast = 262 feet
(Air resistance neglected)
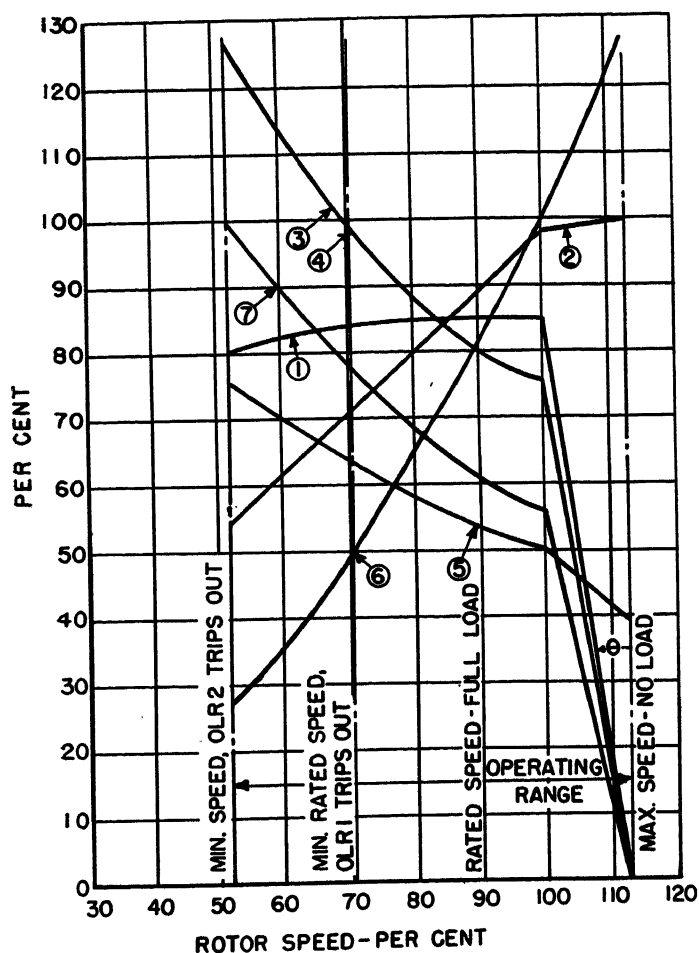
Fig. 7. Performance curves of rotary snowplow fitted with electric drive

1—Power to throw snow, 100%=1,600 hp
2—Generator terminal volts, 100%=862 volts
3—Generator load amperes, 100%=1,800 amperes
4—Motor armature amperes, 100%=900 amperes
5—Motor field amperes, 100%=900 amperes
6—Distance cast, 100%=262 feet ·
7—Rotor torque, 100%=8,000 pound-feet
Note: 100% rotor speed=92 feet per second or
      250 rpm on a 7-foot rotor

Fig. 8. Performance characteristics of stabilizing exciter of electric drive

1—Speed, 100%=1,600 rpm
2—Shunt field amperes, 100%=15 amperes
3—Differential field amperes, 100%=30 amperes
4—Terminal volts, 100%=37.5 volts
5—Armature amperes, 100%=660
6—Power output 100%=25 kw
Note: 100% rotor speed=92 feet per second
      or 250 rpm on a 7-foot rotor

This removes the field of the pusher unit generator, returns its engine to idle, and thus stops the forward motion of the train and gives the plow a chance to clear itself.

If, for any reason, the torque demand of the rotor should exceed the safe maximum of the drive, OLR2 operates. This removes the field of the power unit generator and returns its engine to idle. After a moment's delay, the power contactors open, removing all power from the drive. The train may then be backed up and inspection made for hidden obstructions.

This drive has inherent speed regulation. However, in case a component failure causes the rotor to overspeed, the mechanical overspeed switch will operate. This removes the field from the power unit generator and returns its engine to

idle. After a moment's delay, the power contactors are opened and all power is removed from the drive.

Performance

To illustrate the performance of an electric rotary snowplow drive, consider the one designed to operate with a 1,600-hp standard Diesel-electric power unit. This drive has four GE-731 traction motors (powered by the GT-581 generator in the power unit) and is stabilized by a GT-1185-M1 exciter. The performance of this drive when operating at its rated power in the eighth throttle notch is shown in Fig. 7. From curve 1 it may be seen that 85 per cent (%) of the 1,600-hp input to the main generator is available to throw snow. The other 15% supplies the electric and mechanical drive

losses and the rotor windage loss. This drive is able to load the power plant to its rated capacity from rated plow speed to stalling.

Should the rotor suddenly become unloaded, its speed rises but will not exceed the safe limit of 113% of maximum full-load speed. The slope of the speed-power curve at less than rated loads is designated by the angle θ. This angle is a measure of the speed regulation in this region, and can be varied by the setting of the resistor $R1$. Angle θ is made large to keep current surges to a minimum. The reasoning behind this is as follows. When shock loads are imposed upon the rotor, the energy to supply them comes from two sources: kinetic energy available from the drive as it loses speed, and the surge of electrical power from the main generator. The greater the speed droop, the less the power surge. Since this power surge appears in a constant voltage region, it shows up as a current
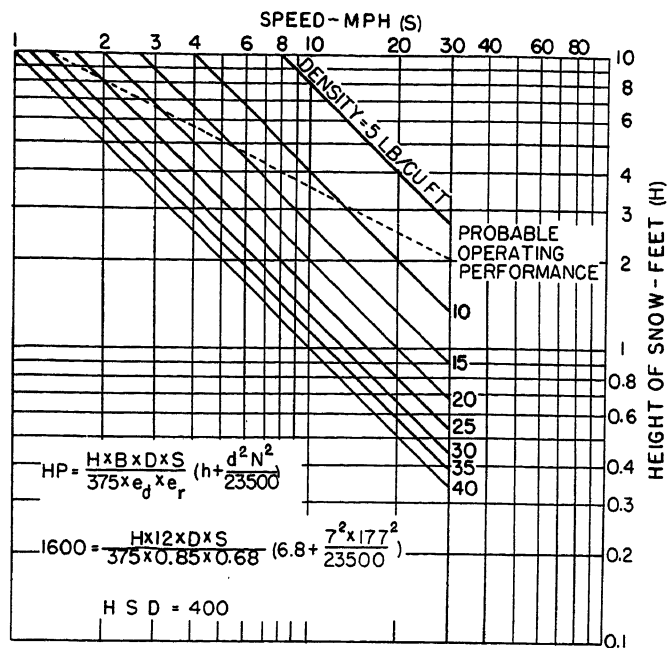
Fig. 9. Performance curves at various snow densities; electric-drive plow powered by a 1,600-hp power unit
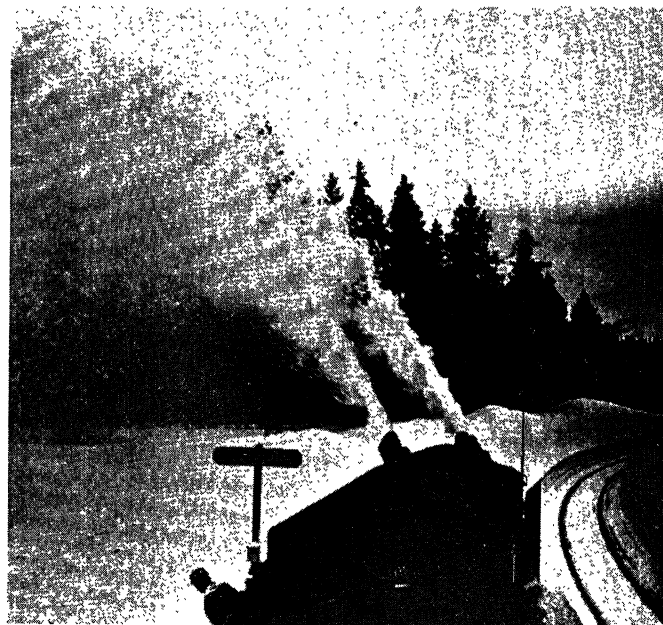
Fig. 10. Electric-drive plow with transverse rotors clearing double track in mountainous terrain

surge in the motors and generators. If the rotary kinetic energy does not come to the aid of the electric equipment, the surge may be so violent that the machines will tend to flash over. Angle θ is purposely made large enough to avoid this.

From the generator volt and ampere curves (2 and 3 in Fig. 7) it can be seen that maximum full-load rotor speed corresponds to the upper corner point of the generator volt-ampere characteristics. Curves 3 and 4 show that $OLR1$ is set to protect the drive from electrical overloads, while $OLR2$ is set to protect it from mechanical overloads. Curve 5 illustrates the continual and gradual automatic adjustment of the strength of the drive motor fields. Curve 6 shows how the distance which the snow is cast varies as the square of the rotor speed. Curve 7 is the speed-torque characteristic of the rotor.

The performance of the stabilizing exciter is shown graphically in Fig. 8. Curve 1 shows that the exciter speed varies directly with the rotor speed. Curves 2 and 3 are plots of the shunt and differential excitation respectively, while curves 4 and 5 depict the resultant variation in exciter volts and amperes. Curve 6 illustrates just how little power is required to

achieve this fine control over so great a speed range. This system responds rapidly to sudden variations in load. The excitation of the drive motor fields was purposely chosen at such a value that it could be held relatively constant across the whole operating speed range, hence the eddy currents set up in the drive motor frames by sudden transients help to maintain rather than destroy system stability.

The depth and density of the snow may vary over wide extremes. The family of curves in Fig. 9 illustrates the effect that depth and density of the snow have on the forward speed of the plow.

To satisfy the last requirement, the plow must be able to track well even under maximum rotor reaction torque. In the type of plow where the axis of rotation is perpendicular to the rails this is no problem. In the type shown in Fig. 2, however, where the axis of rotation is parallel to the rails, the reaction torque in pound feet must never exceed $WG/2$ where W is the weight of the plow in pounds resting on the front truck and G is the track gauge in feet. During normal operation the reaction torque never exceeds 15% of this limiting value. In the event of abnormal operation, for example, the

rotor striking a hidden boulder or a buried telephone pole, the instantaneous peak torque would exceed this 15% considerably, and would undoubtedly shear the pins which are provided to protect the drive. Though the speed of response of this drive is rapid, the rotary kinetic energy will shear the pins rather than wreck the rotor under unusual circumstances.

Conclusions

The drive described in this paper makes more power available to the rotor over a broader speed range than any other known. It features low initial and maintenance cost, high reliability, over-all simplicity, standard interchangeable components, inherent stabilization, and extreme adaptability. It should prove a great aid in overcoming the difficulties of winter railroading in mountainous areas.

References

1. Snow Removal in Mountain Districts, L. Wylie. *Railway Age*, New York, N. Y., vol. 133, Dec. 15, 1952, pp. 57–59.

2. The Electric Rotary Snow Plows of the Swiss Railways, E. Anderegg. *Bulletin Oerlikon No. 281*, Zurich, Switzerland, pp. 1982–90.

---◆---

Discussion

W. S. H. Hamilton (65 Edgewood Avenue, Larchmont, N. Y.):
I first became interested in rotary snowplows, particularly those of the type shown in Fig. 2, on a trip in the Bitter Root Mountains on a rotary

snowplow (or rather, four of them) in March 1917. We spent 4 days going 30 miles, but were not by any means in motion all the time because of various accidents that occurred which are distinctive to this kind of work, especially in mountainous territory.

A rotary snowplow has many different

and difficult functions to perform in operation, some of which are not brought out in the paper.

1. The wheel must be reversible so that snow can be thrown to either side of the track. The deflecting hood which is shown just above the wheel casing itself

in Fig. 2, also has to be reversed at the same time.

2. This reversal of direction of throw of the snow is required mainly to enable the snow to be thrown to the downhill side of the track when working on a hillside (which all mountains seem to be plentifully provided with), but also to avoid throwing it on buildings that may be along the right of way. For this reason it is often necessary to speed the plow up at times in order to throw snow over and beyond such things as buildings, passing sidings, etc. It must be kept in mind that in the stream of snow there are often small rocks, etc., which can do considerable damage if they strike anything. For these reasons it is customary to protect the windows of all railroad buildings on the downhill side of the track with heavy boarding during the winter season. It can be easily imagined that a rotary snowplow would not be popular in congested areas, such as those around most eastern cities.

3. In operation it is important that the wheel speed be kept high enough and the forward speed of the plow low enough so that the wheel will free itself and not clog up and stall. This requires considerable judgment on the part of the operator, especially when the plow is passing through snow of varying depths.

4. The wheel must not race when it comes out of the snow momentarily as often happens in passing through drifts or snow of uneven depth. In the plows driven by steam engine, a man is stationed at the throttle of the steam engine which drives the wheel in order to catch any tendency to race.

5. It is important that the plow be pushed forward at a steady speed. In the days of steam locomotives, this was best accomplished by using a Mallet locomotive which has a high weight on drivers and whose speed is relatively unaffected by changes in pushing power required. This forward speed may vary from 2 to 4 mph in deep snow to 15 mph in light snow, or where the rotary is used to clean up tracks that have already been plowed out.

6. It is very important that the wheel be kept turning at a stable speed of about 50 to 60 rpm, when the plow has been pulled back from actual snow-cutting, as is often required on account of obstacles other than snow on the track, in order to prevent the wheel from freezing to the casing of the plow. While these plows will dispose of trees up to approximately 4 inches in diameter and fair-sized rocks, larger obstacles are often encountered, which must be cut away by hand before plowing can be resumed.

If memory serves, most of the rotary snowplows were built before 1910, with the exception of a very few built within the last few years. While the author gives some good equations for the performance of the wheel, etc., I suspect that the designs were actually evolved by the trial-and-error method.

Likewise, the connections immediately preceding those of Fig. 4 were evolved through very extensive calculations, which showed that four armatures in parallel with four fields in series would operate satisfactorily. These proved very much in error in actual test, and Fig. 4 is, of course, an improvement over what it was possible to work out rather hastily in the field.

I believe the arrangement proposed by the authors will satisfactorily meet all the requirements, with a minimum of attention demanded from the plow operator. Flying snow affects visibility from the cab of a rotary which means that the operator must either keep his head out the side door or use a clear-vision window which will open in front of him. Neither arrangement is very pleasant from the standpoint of comfort. I hope that the arrangement proposed in this paper will prove so attractive to the railroads of this country that all the rotary snowplows now in use will be electrified.

J. L. Partridge (New York Central System, New York, N. Y.): I should like to relate a little of the history pertaining to the development of the design of the Diesel-electric rotary plow.

In 1950 the New York Central Railroad decided to convert one of their steam rotaries, built in 1889, to electric drive to operate with a 1,600 hp Diesel electric locomotive. The Burlington Railroad had just remodeled one and it was decided to pattern the New York Central plow after this mechanically. The arrangement is the same as shown in Fig. 3.

The Burlington's electrical scheme used four traction motor armatures in series-parallel connected across the main generator of the power locomotive. The four motor fields were kept separate from this high-voltage circuit, being connected in series across the output of the power locomotive's 74-volt auxiliary generator. Another generator geared to one of the plow's drive shafts was also connected in parallel with the auxiliary generator to stabilize rotor speed at no load.

The first plan used on the New York Central's plow was to wire the four motor armatures in parallel connecting the positive side of each to the positive terminal of the power locomotive's main generator. The negative motor armature leads were connected to one end of the four motor fields all wired in series. The opposite end of the fields was then joined to the main generator negative terminal, thus putting the four parallel armatures in series with their fields which were in series.

An exciter generator was connected in parallel with the four fields to provide additional field current at no load to limit the rotor's tendency to overspeed. This generator was geared directly to one of the power shafts of the plow.

The plan was calculated to maintain excellent speed regulation between no load and full load, the maximum rotor speed change being in the neighborhood of 14% or a variation between no load and full load of 172 to 150 rpm.

The motor armatures, unlike Diesel locomotive motors, were tied together mechanically so that relative slippage between them was impossible, since they were all geared to one rotor wheel. With the four fields connected in series, the motors all had identical field currents and, since the armatures were in parallel their armature voltages were supposed to be the same. However, because of slight varia-tions in back electromotive force, an un-balanced condition was created which set up a violent bucking action between motors. It was so pronounced that at no load (rotor running free) with the 8-notch controller in the number 2 position the motors bogged down completely.

The plow was rewired as a result, using a scheme very similar to that shown in Fig. 4, except that a constant source of excitation was still supplied to the exciter field which, however, was made to vary with changes in the controller setting. No mechanical changes were made to the plow.

With two sets of motor armatures in series placed across the main generator, it was necessary to reduce the field excitation to maintain the same no-load speed. This meant operating much lower on the satura-tion curve, hence there was more speed variation with a variation in field current. And since the fields were wired to conduct the full-load current that passed through their respective armatures (assuming none was shunted through the exciter), it could be expected that field current would increase markedly with increased load.

The chief reason for forming the two main parallel paths through the motor armatures and their respective fields was to ensure that any attempt of one motor to take on more than its share of the load would be telegraphed to its field winding, thereby tending to boost its back electro-motive force and thus reduce the motor's taking on more than its due. However, in order to make this effective, some re-sistance had to be introduced between the positive side of the exciter armature and each pair of motor fields. It was finally decided that 0.03 ohm in each leg was the minimum value that should be used to maintain a stable system. These corre-spond to R2 and R3 in Fig. 4.

This meant that total resistance in each leg of the exciter circuit was 0.03 ohm plus 0.02 ohm for the motor fields, whereas originally there was one path of 0.04 ohm which consisted entirely of motor field resistance. Therefore, fluctuations in motor field current would not now have as much effect on exciter output under the new scheme as they did originally since the 0.03-ohm resistance does not carry any of the motor load unless exciter current is re-versed.

In the original scheme an increase in load tended to boost motor field current. But this increase had a strong effect on the exciter and caused its output to drop rapidly, thus resulting in a relatively small net increase in motor field strength and a correspondingly small effect on rotor speed. In the second scheme the exciter current was much less affected by increased loads and thus the motor field strength built up with the additional motor armature current while exciter current decreased little.

In the first test in snow in December 1951, at Watertown, N. Y., using this wiring scheme the no-load rotor speed was set at 150 rpm. When plowing in hard-packed snow to a depth of 4 to 5 feet with the wings of the rotor housing extended out to the side, the rotor wheel slowed to about 105 rpm and track speed had to be reduced to about 2 miles per hour. Hp input to the motors at this point was only about 400.

To further decrease excitation of the motor fields under load a relay was installed whose operating coil is sensitive to increases in motor armature current and whose contacts cut in additional resistance in series with the exciter field when the relay picks up. This has not been tested to date because of a lack of severe snow conditions. The coil of the relay was operating during the 1951 test but the contacts were not connected. The relay picked up at about 25% of full load.

As brought out in the paper, it is necessary to maintain a relatively high motor speed in deep snow if normal track speed is to be maintained. To do this, motor field current must not increase too sharply when full load is applied. Therefore, some of the load current must be shunted through the exciter armature in the reverse direction to normal flow or, in other words, cause the exciter to operate as a motor. It is this requirement that led to the adoption of the differential field as described in the paper. Such an arrangement provides continuous, smooth control of rotor speed and is much

preferred to the intermittent action of a relay.

A few other points about the New York Central plow may be of interest. When the electric drive was applied, the original rotor and underframe built in 1889 were reused with some strengthening of the frame added.

A hostler's controller was installed in the plow cab to control the Diesel engine speed in the power unit coupled directly behind the plow. No electrical connections were made to the pushing unit except for the installation of sound-powered telephone equipment to enable the conductor stationed in the plow cab to communicate with the engineer of the pusher locomotive. The plow operator could also talk to the engineer through his own chest microphone and head set.

Originally the plow was operated with a 1,600-hp freight locomotive but was later altered to operate with a 1,600-hp road-switching locomotive which is the most prevalent type of power in the Watertown area.

A. H. Hoffer and R. E. Willhite: We thank Mr. Hamilton and Mr. Partridge for their interesting history of the electric-drive rotary snowplow as applied to the operation of the New York Central Railroad. No attempt was made in our paper to give the history of electrically driven snowplows. Investigation revealed that the few in existence today were developed under circumstances peculiar to the railroad which built them.

Recent tests on the drive described in the paper substantiate the predicted speed power performance graph. A technique has been developed which eliminates the connections between the exciter shunt field and the auxiliary generator on the power unit. Instead, half of the exciter shunt field is connected across one set of the drive motor field and the other half of the exciter shunt field is connected across the other pair of drive motor fields. Consequently, this offers the advantage of requiring only two power leads instead of four between the power unit and the snow plow.

Residential Electric Space Heating in Detroit for 1952–1953 Heating Season

A. E. BUSH
ASSOCIATE MEMBER AIEE

R. P. WOODWARD
NONMEMBER AIEE

Synopsis: In this paper the load characteristics of 29 domestic electric space heating installations in Detroit during the 1952-1953 heating season are given. The theoretical undercapacity of most of these installations are pointed out as well as the ineffectiveness of using the measurement of kilowatt-hours (kw-hr) per degree day per 1,000 cubic feet of heated space when comparing different houses.

THERE has been a slow but steady growth in the number of homes with electric radiant panel heating in The Detroit Edison Company service area since 1949. At the present time, 135 residences are being heated electrically as well as some 167 dwelling units in 17 apartment buildings and two 12-room motels. This growth in number of installations is not phenomenal compared with the increase in other electric appliances such as television. The potential magnitude of this load with even a very low saturation, however, is great enough to invite further investigation.

The policy at Detroit Edison has been to obtain as much data as possible on the load characteristics of homes with this type of heating. Only in this manner can such problems as distribution design,

desirability of promotion, and adequacy of current rates be studied and a solution found that is economically acceptable to the electrical utility.

Prior to the 1952-1953 heating season, data could not be obtained for an entire heating season from more than 20 customers with electric house heating. The principal reason for this was the high cost of operation compared to gas or oil, and the subsequent use of such fuels either partially or entirely. Many of the early installations were in homes previously heated by other methods and, in some cases, not adequately insulated. The high electric bills that resulted caused many of these people to change back to their former heat source.

There has been a definite trend in the past year or two to installations in new homes designed and insulated for this type of heating. Adequate demand and use data on 29 homes were obtained during the 1952–1953 heating season. This number is large enough to allow for the diversity between customers and the demands should be indicative of those expected from a much larger group. It should be pointed out that the group metered is neither a random nor a selected

sample. Meters were installed on all electrically heated homes which appeared likely to continue with this type of heat throughout the winter period. In some instances there was insufficient space to install metering equipment, or the customer was definitely against the installation of additional meters, or it was not possible to separate the heating circuits for metering. The results cannot, therefore, be interpreted as being a scientifically selected sample representative of the entire population of present house-heating customers in the Detroit area. They are perhaps a good representation of the types of homes which will retain this method of heating.

For these tests, type G-9 integrating 30-minute demand meters were installed for the winter months on the house-heating load only in each home. Record kw-hr meters also were placed on the space-heating load only and the total use for the home was read from the billing meters. The heating kw-hr use figures were not available for the complete year in 12 cases, so the missing figures were estimated from available data. A record of actual inside temperatures maintained was not obtained.

One further factor should be considered in interpreting this report. The 1952–

1953 heating season was abnormally warm. The degree days from September 1952 through June 1953 totalled 5,849 as compared with the normal for Detroit of 6,396 based on the 30-year period from 1921–1950 inclusive. It is interesting to note that this normal value has not been reached in Detroit for the last five heating seasons. The minimum weekday temperature during the test period was 11 degrees Fahrenheit (F) on January 7, 1953 as compared with a more normal minimum of −5. Estimates of the probable effect of the colder weather on the space-heating use and demands are included later in the paper.

Customer Characteristics

The electrically heated homes in this study ranged in size from three to six rooms with an average of five rooms. Nineteen of the 29 were fully insulated while 10 had insulation in walls and ceilings but not in the floors. All but six of the homes were completely equipped with storm sashes on windows and doors.

The average floor area was 974 square feet with a range from 494 to 2,026 square feet. The volume of heated space (inside measurements) also exhibits a wide variation from 3,830 to 16,381 cubic feet, with an average of 7,644 cubic feet. The average connected load of the heating units was 8.83 kw, the smallest being 5.0 kw and the largest 22.75 kw. Based on these figures, the average installation has a theoretical output of 30,137 Btu per hour (connected load X 3,413 Btu per kw-hr). This is only 72 per cent of the average calculated heat loss of 41,785 Btu per hour based on an outdoor temperature of −10 F and an indoor temperature of 70 F.

In only three of these homes does the theoretical output exceed the calculated heat loss and five more are between 80 and 100 per cent. This apparent shortage of capacity has not been critical since the minimum design temperature of −10 F has not been reached in Detroit since electric panels have been in use. The lowest value has been −4 F in 1951. In a continued cold spell with outdoor temperatures near zero degrees for several days, it is probable that many of these electrically heated homes will be unable to maintain a satisfactory indoor temperature.

Load Characteristics

The energy use and demand figures are summarized in Table I. Three separate groups of figures are included: a full year, the coldest weekday (day with the greatest number of degree days), and the average weekday of the coldest week. Annual energy use for space heating averaged 11,412 kw-hr of a total of 14,560 kw-hr per customer. The other use of 3,148 kw-hr indicates that customers with electric heating own more of the other electric appliances than the average home. Those in the sample used almost half again as much electricity for other than house heating as the average Detroit Edison domestic customer. These figures are exclusive of the amount used for unmetered off-peak water heating.

This emphasizes the importance of collecting load data regarding electric space heating. Even these relatively high-use customers use almost four times as much electricity for their heating in a mild winter as they do for lighting and appliances. In a normal heating season there would be an increase in energy use for space heating to about 12,500 kw-hr per year. A 16-per-cent saturation of electrically heated homes would double Detroit Edison's average annual metered sales to farm and residence customers.

One of the factors frequently used in comparing electric heating installations has been kw-hr per degree-day per 1,000 cubic feet of heated space. The average value for the 29 homes in this test based on inside measurements was 0.26 with a range from 0.12 to 0.48. The wide spread in these ratios indicates either that variables are included that make comparisons impracticable or else that there are great differences in living habits, insulation, or indoor temperatures maintained.

There is actually little reason to expect this ratio to be a constant for homes in any one locality and even less when comparing homes in different cities with different climatic conditions. Certainly, a house built in Detroit for electric space heating will be much better insulated and, consequently, have much lower heat losses than a home of similar size in Tennessee. It seems rather useless to make comparisons based on this ratio or to attempt to establish a normal value. Even when all of these installations are in the Detroit area, this ratio for individual customers shows practically no tendency to center around any particular value. They are quite evenly distributed between the highest and lowest values so a normal or frequently occurring value does not exist. This has been found to be generally true in other localities where house-heating data are available.

A more meaningful ratio which can be used is kw-hr per degree-day divided by 1,000 Btu hourly heat loss per degree of designed temperature differential. This is a restatement of the equation prepared by the Federal Housing Administration for calculating the electric energy for space heating and should yield a value of

Table I. Load Characteristics of Electric Radiant Panel Heating in Detroit for the 1952–1953 Heating Season

Number in Test, 29; Average Number of Rooms, 5.0; Connected load, 8.83 Kw

	Test Period		
	Annual Period 6–1–52 to 5–31–53*	Coldest Weekday 1–6–53†	Average Weekday (Coldest Week) 1–5–53 to 1–9–53
Energy use per customer, kw-hr			
Space heating	11,412	86.9	85.2
Other use (excluding off-peak water heating)	3,148		
Total	14,560		
Degree days, 65 F base	5,834	49	42
Space-heating use, unit figures			
Kw-hr per degree-day per customer	1.96	1.77	2.03
Kw-hr per degree-day per 1,000 cubic feet	0.26	0.24	0.27
Kw-hr per degree-day÷1,000 Btu heat loss per degree of designed temperature differential	3.76	3.39	3.89
Coincident maximum demand (30-minute integrated)			
Kw per customer	4.77	4.63	4.22
Day	Wednesday	Tuesday	
Date	1–7–53	1–6–53	
Time (half-hour ending), AM	9:00	8:00	9.00
Outdoor temperature, F	12	12	22.6
Noncoincident maximum demand, kw (per customer)	7.18	6.04	6.42
Load factor			
Based on coincident demand, per cent	27.3	78.2	84.1
Based on noncoincident demand, per cent	18.1	59.9	55.3
Coincidence factor, per cent	66.4	76.7	65.7
Demand factor‡	81.3		

* The energy use data were available for 17 of the 29 customers for the full year period. Data for remaining 12 customers were expanded to an annual basis. Coincident demand data were available for all customers for the weekdays, January 5 to 9 inclusive.
† Data are shown for coldest day of coldest week (January 6, 1953). High use and demand day was January 7, 1953.
‡ Ratio of noncoincident maximum demand to connected load.
Note: All values except annual energy use per customer pertain to space-heating load only.

5 for radiant panels according to their estimates.

The average of the 29 Detroit installations was 3.76 with a range from 1.36 to 7.12. There is a greater tendency for these values to group with the exception of one very high and one very low instance. The deviations that are still present are undoubtedly caused by differences in the number of persons at home during the day, the presence of small children or infants, the number of rooms that are kept heated, and the temperature level maintained.

Using this ratio to compare electric space heating in various localities brings out the fact that the electric rate or warmer temperatures have a definite effect on usage. The value of this ratio for most homes where the electric heating rate is above 2 cents per kw-hr is between 2 and 6 with an average close to 4. Where the cost of heating is closer to 1.5 cents per kw-hr, the annual use adjusted by heat loss and degree-days appears to range from 4 to 8 with an average of about 6.5. These lower cost areas are in warmer locations so this difference in corrected usage may be caused by one or both of two factors: with lower rates and smaller bills the tendency is lessened to practice economies in the use of electric heating such as turning down the thermostats in little-used rooms or at night; and accustomed to warmer temperatures generally, there is an inclination to keep the thermostats at higher settings to provide comfort to the occupants. Whether one or both of these factors are the cause, it does seem definite that even this ratio has different values in different areas. While a factor of 5 is probably close to the average for all house-heating areas, 4 would be more logical for the northern localities with higher rates, and 6.5 more representative of the southern communities with lower kw-hr charges.

A single ratio such as this cannot be expected to be true for any single home in any area; however, it does give better comparative values than the so-called heat factor (kw-hr per degree-day per 1,000 cubic feet of heated space). The variation in living habits and temperatures maintained cause substantial deviations which cannot be adjusted by these simple relationships. Table II is a tabulation of data for 29 individual space-heating installations, arranged in order of space-heating kw-hr use per degree-day per 1,000 Btu heat loss per degree-difference in design temperature. The table includes size of home, number of persons, daytime temperatures maintained, and similar customer characteristics. It must be emphasized that all data including temperatures were obtained from customer interviews and may not, therefore, be completely reliable.

The various factors affecting space-heating use are interrelated to such an extent that direct correlation of any individual factor with use does not appear warranted. With the degree of insulation largely eliminated through the adjustment for heat loss, the most significant difference between customers with high and low factors was the date of original installation of space-heating panels. All of the customers with ratios in excess of 5.0 had installed electric space heating just prior to the 1952–53 heating season. Customers with longer experience appear to have decreased their use of electricity by: supplemental heating by fire places or space heating with wood, coal, gas, or oil; adaption of living habits to lower temperature levels; and more consistent reduction of temperature at night and during absences from home.

The maximum 30-minute coincident demand of the 29 test customers was 4.77 kw, reached on Wednesday, January 7, 1953 from 8:30 to 9:00 AM. The minimum weekday temperature for the 1952–1953 heating season of 11 F occurred the same morning although the coldest day based on the number of degree-days was January 6. The average noncoincident maximum demand was 7.18 kw. Annual load factors were 27.3 per cent and 18.1 per cent based on the coincident and noncoincident demands respectively. These load factors are higher than had been found for the electrically heated homes in the three previous heating seasons.

It is probable that the average maximum demands, both coincident and non-

Table II. Individual Customer Characteristics Affecting Use of Electricity for Space Heating

Customer	Space-Heating Factor*	Heated Space, 1,000 Cubic Feet	No. of Rooms	No. in Family	Daytime Temperature, F			Temperature Reduced Nights	Year Installed	Other Factors‡
					Living Room	Kitchen	Bedroom			
A	1.36	10.2	5	4†	72	70	70		1950	supplemental heat
B	2.48	6.2	4½	2	72	70	65	x	1949	heated basement
C	2.56	4.4	5	4	62	62	62	x	1950	not occupied days
D	2.80	6.8	6	3	72	72	72	x	1949	oven heat in kitchen
E	3.04	8.4	6	5†	70	70	50		1949	heated basement
F	3.04	7.9	4½	3†	70	70	65		1952	some supplemental heat
G	3.12	11.1	5	2	72	72	72		1949	apartment over store§
H	3.12	4.5	4	2	72	70	70	x	1950	heated basement
I	3.20	16.4	6	not available					1952	
J	3.20	7.9	4½	2	70	65	65	x	1950	heated basement
K	3.28	11.3	6	2	72	72	60	x	1951	
L	3.36	10.7	6	not available				x	1952	
M	3.36	6.0	4	3†	75	68	70	x	1950	some supplemental heat
N	3.52	11.6	6	4	72	72	50	x	1952	
O	3.84	6.1	5	3†	72	72	72	x	1950	
P	3.84	10.1	5	2	70	70	70	x	1950	
Q	3.92	5.2	5	4†	70	70	70	x	1949	heated basement
R	4.00	5.6	5½	4†	75	75	75	x	1950	some supplemental heat
S	4.32	5.1	6	5†	72	72	70	x	1950	
T	4.40	5.6	4	4	72	65	70	x	1952	
U	4.48	5.7	4	4†	70	70	70	x	1950	
V	4.72	11.9	6	5†	68	68	65	x	1951	some supplemental heat
W	4.80	6.0	5	2	70	70	60	x	1950	heated basement
X	5.04	5.0	4	4	70	70	60	x	1952	
Y	5.02	8.8	5	3†	not available			x	1952	
Z	5.12	7.1	7	5†	73	73	73		1952	
AA	5.44	5.4	5	2	75	70	50		1952	
BB	5.52	6.7	4	5†	70	70	70	x	1952	
CC	7.12	3.8	3	3†	70	70	70		1952	some supplemental heat

* Space-heating kw-hr use per degree day per 1,000 Btu hourly heat loss per degree difference in indoor and outdoor design temperature (80 F).
† Indicates families having children.
‡ Supplemental heating includes fireplaces, oil, coal and gas stoves, primarily in basement locations.
§ All other installations were in detached single dwellings.

coincident, would have been higher had the design temperature of −10 F been reached during the test period. Based on data available for the past four heating seasons, it appears that the coincident maximum demand would be about 6.4 kw with this lower temperature. Assuming a normal season heating use of 12,500 kw-hr, the annual load factor based on the coincident maximum demand would be about 22 per cent.

Daily load curves of the average of the 29 customers for January 6, 1953 (the day with the greatest number of degree-days) and January 7, 1953 (the day with the minimum temperature) are shown on Fig. 1. On January 6, there were 49 degree-days, the average use per customer was 86.9 kw-hr and the maximum coincident demand was 4.63 kw. There were 47 degree-days on January 7 with 90.8 kw-hr used per customer and a maximum demand of 4.77 kw. The hourly outdoor temperatures for each day are also plotted on the same graph.

Although the general shapes of the two curves are very similar, there are differences which are interesting to note. There appears to be a direct relationship between temperature and average demands for these 29 customers when comparing these two days. Before 8 AM, when the January 7 temperature was lower than the temperature for the previous day, the average demand was also lower. After 8 AM the temperatures were reversed with January 7 being the warmer of the 2 days. The demands after 8 AM were also higher on January 7.

This apparent inconsistency probably is accounted for by climatological factors other than the temperature. From mid-
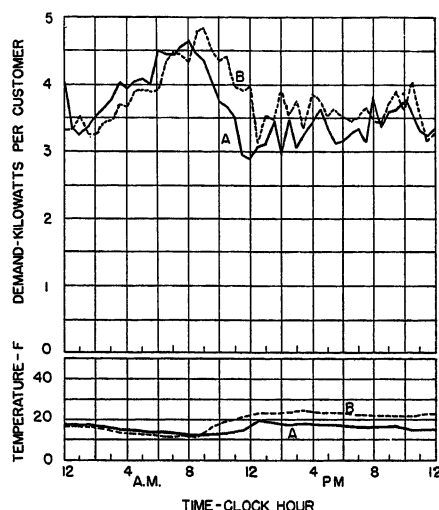
Fig. 1. Detroit space-heating coincident load curves and temperatures for 1952–1953 heating season

A—Coldest weekday, Tuesday, January 6, 1953

B—High use and demand day, Wednesday, January 7, 1953

night to 8 AM the wind was 13 to 14 miles per hour on January 6 and less than 7 miles per hour on January 7. The demands were greater on the morning with the stronger winds. During the daylight hours, the wind speeds were very similar but there was a difference in sunshine. On January 6 there were over 5 hours of sunshine as compared with less than 2 hours on the following day. The demands reflect this difference with the lowest demands on the day with the most sunshine.

This comparison emphasizes once again that house-heating use is not dependent

upon temperature or degree-days alone. While this is probably the most important variable, such things as cost of electricity, insulation, wind velocity, amount of sunshine, living habits, etc., have enough influence so that exact comparisons of temperature and kw-hr usage cannot be expected.

Conclusions

The cost to the customer of electric space heating at current residential rates, even with the small and well-insulated home, is in the neighborhood of $275 per season. Some of the larger homes may run as high as $600. This does not compare favorably with the present costs for other forms of heating. Despite this, many average income families are interested in electric heat because the cost difference is a small part of their yearly income. This, coupled with the advantages of using electric heating, tends to minimize the differential in operating costs.

From the utility standpoint, the cost to serve this type of load with a high demand occurring during the winter period and a load factor lower than that of the present residence customer may well be excessive when compared with other domestic uses. On the other hand, the application of electric heating has only begun and no one has the complete answer. With widespread growth in the use of air-conditioning, it is possible that electric heating may become an offsetting load. Because of its susceptibility to a wide variation in control, it is possible that electric heating in itself may be improved with respect to load factor.

Discussion

John C. Beckett (Wesix Electric Heater Company, San Francisco, Calif.): The paper by Mr. Bush and Mr. Woodward is of significant value to the general knowledge regarding the load characteristics of homes heated with the resistance-type electric heating equipment. The authors have been very careful to point out that the heat factors frequently used in comparing electric heating installations, such as the kw-hr per degree-day per 1,000 cubic feet of heated space, vary over a wide range. They emphasize, too, that this is to be expected when comparing homes in different cities with different climatical conditions. It might be well to add that this variation occurs not only because of different climatical conditions and different living habits of the occupants but most particularly because of different types of heating equipment. This test involved only one type of equipment and thus is

not comparable to similar tests made in other areas where different equipment was used.

In an unpublished AIEE paper by Sharp and Opdenweyer individual room radiant convection heaters were compared with central electric furnaces. A statistically significant number of houses was tested in each case; the houses were of similar size, construction, and insulation. The average difference in energy consumption was 25 per cent greater for the central systems. A recent survey of a similar nature[1] revealed an even greater difference involving a considerably larger number of homes. In this case the central systems under comparable conditions consumed 60 per cent more energy.

The authors point out that a normal or most frequently occurring value does not exist and house heating use is not dependent upon temperature or degree-days alone; thus, the equations that attempt to reduce load estimates to a single simple relationship are of little value. The authors

suggest that a more meaningful ratio might be the kw-hr per degree-day divided by 1,000 Btu hourly heat loss per degree of design temperature differential, and that this is the basis of the proposed Federal Housing Administration equation for calculating energy for space heating. I do not follow the authors' reasoning here as it would appear to me that this ratio or relationship is no more consistent than others now used which take into account only heat loss and degree-days. The so-called heat factor, criticized by the authors and involving the relationship of kw-hr per degree-day per 1,000 cubic feet of heated space, is perhaps better in that a schedule of factors is applied depending upon the heat-loss characteristics of the house, the locality of the house geographically, and the type of equipment being used. The present heat-factor equation does have the rather serious defect of assuming heat loss to be proportional to volume, but this could be corrected rather easily.

Mr. Homer Bender, the originator of the

heat factor, did not intend that one set of factors would be applicable throughout the country, but rather that heat factors for various conditions must be developed and applied. As far as the electric space-heating equipment manufacturers are concerned, it has been the experience of the ones making heaters for 20 years or more that a reasonable estimate of energy consumption can be made where the size of the home as well as its construction, geographical location, and type of equipment are all taken into account.

Three significant fundamental variations are taken into account in using a schedule of heat factors; not all of these are discussed by the authors. The Federal Housing Administrations equation takes into account only the calculated heat loss of the structure without any regard to its geographical location or equipment used.

Undoubtedly, a schedule of heat factors should be developed that would be useful to the industry as a whole. For example, one set of values might be applicable to the West Coast area including the states of Washington, Oregon, and California; another set for the inland Northwestern States, extending from eastern Washington to the Great Lakes and as far south as southern Utah and Colorado; another area might include the Southwestern States; another the Southeastern States; and still another the Eastern States. The difference in energy-consumption characteristics of the different types of heating systems now available—radiant heating cable in the floor, radiant heating cable in the ceiling, individual room radiant convection wall heaters (including baseboards), individual forced air-circulation heaters, central warm air systems, and central hot water systems—are quite well known. Load characteristics have been published by the Tennessee Valley Authority.[2] The references cited and the present paper effectively establish the characteristics of at least some of these systems.

One of the principal reasons for the variation in heat factors for different localities has not been adequately understood, and this should be of technical interest to all concerned with the load characteristics of heating systems. This is the fact that the degree-day does not give a true indication of electric heat requirements. The definition of the degree-day is the number of degrees that the average daily temperature is below the 65 F base. This average temperature of maximum and minimum is a poor indicator for many areas as to what the heating requirements might be. For example, California climate is characterized by warm days and chilly nights. The hourly temperature during the milder months of fall and spring will take a momentary drop to a relatively low level, but never remains there very long and, during the day, will be consistently in the 60's. Under these conditions the degree-day falsely indicates a greater need for heat than is really required. Added to this, people are less demanding of heat at 4:00 AM than at 5:00 PM. What is needed for more accurate determination of energy requirements is the degree-hour. An article by Edwin Fleischmann emphasizes this point.[3] A compromise solution might be to redefine the degree-day as the difference between the mean hourly temperature (taken each hour over a 24-hour period) from a 65-degree base instead of an average of the maximum and minimum temperatures only. The difference in hourly temperatures is not the only variation which affects energy consumption but it does appear to be the most significant one; others that might be mentioned include wind and relative humidity.

REFERENCES

1. SHORT METHOD FOR ESTIMATING ELECTRIC HOUSE HEATING LOAD, Robert E. Sinclair. *Air Conditioning, Heating and Ventilating*, Chicago, Ill., Jan. 1955.

2. GUIDE FOR CALCULATIONS OF ELECTRIC SPACE HEATING. Tennessee Valley Authority, Chattanooga Tenn., revised, Sept. 1947, p. 11.

3. HOW OUTDOOR TEMPERATURE AFFECTS ELECTRIC SYSTEM LOADS, Edwin Fleischmann. *Electrical West*, San Francisco, Calif., June 1954.

A. E. Bush and R. P. Woodward: Mr. Beckett's point that relationships used to estimate electric space-heating kw-hr use are not consistent when comparing individual room heaters with central electric furnaces is well taken. There is also no question that the reliability of the heat factor will be increased when used in conjunction with information regarding the type of heating equipment in a particular home, its construction, or geographical location. This substantiates our point that any one factor of itself does not allow for the many possible variations.

Use of degree-hours in place of degree-days as a basis for heating requirements may be of some assistance. It still would not affect the deviations found in the Detroit data or that for any other one locality as it would simply mean dividing all of the kw-hr figures by a different constant and would in no way adjust for other influences.

The compromise solution suggested by Mr. Beckett would be of doubtful value in the Detroit area although perhaps of some use in areas having different climatical conditions. Comparison of monthly degree-days for Detroit from September 1953 through February 1954 by the two different methods are shown in Table III. These differences do not seem great enough to add any measure of reliability to ratios utilizing them. The ready availability and widespread use of degree-days based on the average of the daily maximum and minimum temperatures would appear of more importance than the possibly increased precision obtained using the average of the hourly temperatures.

Table III. Comparison of Monthly Degree-Days

| | | Degree-Days with 65 F Base | |
		Using Average of Hourly Temperatures Each Day	Using Average of Maximum and Minimum Daily Temperatures
1953	September	99	102
	October	293	264
	November	626	609
	December	979	964
1954	January	1,200	1,200
	February	885	874
	Total	4,082	4,013

The New Look in Lamp Bases

J. O. GEISSBUHLER
NONMEMBER AIEE

THERE are many uses for aluminum in the lamp industry. For example, it is the reflecting coating on all lamps of the sealed beam type; it is used in some lamps as a powder for its beneficial effect on lamp performance, and in the manufacture of a heat-reflecting disk to reduce the base and socket temperatures in some lamp designs. But by far the largest amount of the material is employed in its newest use, the manufacture of bases. This paper is a discussion of data and tests on aluminum as it performs as a base-making material.

Bases are the units that provide the mechanical and electrical connection for lamps. They consist usually of a shell and one or more center contacts insulated from each other and the shell by a glass or a plastic material. Samples of the more common type of base using aluminum are illustrated in Fig. 1. In the fluorescent lamp, the aluminum shell is not used to complete the electrical circuit, nor does it contact a metal socket in any way. No special problems are involved here. In the incandescent lamp base designs, the aluminum shell is the means for mechanically holding the lamp in the socket and also for making the electrical connection.

Traditionally, brass has been the material for use in base manufacture. Why, then, was aluminum decided upon? Twelve to 15 years ago, during World War II, a serious shortage of brass developed, necessitating either a substitution for brass or a curtailment of production. The costly experience of scrapping lamps at the end of the war because of unsatisfactory substitutions led to the search for a more suitable material. The goal was to find a material which would give satisfactory lamp performance, be

in free supply, and the use of which would result in lower costs. In the interval between World War II and the Korean incident, this development was started. It became apparent very quickly that aluminum answered some of the primary requirements. It had good electrical conductivity, mechanical strength, and stability. Cost analysis indicated that it would lower base costs and be in freer supply, as compared to brass. But could it be adapted to the manufacturing methods of the Company and would it perform satisfactorily in service? These were the primary questions to be answered before aluminum could be considered an acceptable base material. The Korean incident, of course, greatly accelerated its development. Brass, again, came in short supply, but by this time those working on the program had some manufacturing knowledge and experience, and had done some field testing of the aluminum base. The results of both the limited manufacturing and field tests were favorable. On this basis production was started.

Over the years, the Company has worked very closely with the manufacturers of brass, to obtain brass strip with a high order of dimensional and physical uniformity. With this strip it has been possible to develop high-speed manufacturing techniques. High-production automatic equipment for making bases and automatic base-filling and base-soldering machines were in use. With

the change of material, many of the difficulties which had been overcome in connection with brass had to be tackled again for the new material.

Here are some of the problems to be solved from a manufacturing standpoint. In base manufacturing, the glass insulation is pressed into the shell at the melting temperature of the glass. Early aluminum alloys annealed at this temperature and were soft, and not very suitable for further handling. This problem was solved by a combination of several factors. The glassing-in temperature of the glass was reduced as much as possible. Aluminum alloy of higher temperature was used and, in the glassing-in operation, the aluminum metal was kept as cool as possible.

In lamp assembly, other problems developed. Side-soldering on automatic equipment here is an essential. On brass, ordinary lead-tin solders made an excellent connection and could be soldered with very low shrinkage. However, lead-tin solders were not suitable for use on aluminum for, in the humidity test, corrosion developed very quickly between the solder and the base material, resulting in a no-contact joint. Cadmium solders and new fluxes were required to make an acceptable side-solder connection. This required new techniques on the finishing machine and resulted in somewhat higher shrinkage. About this time, cadmium became scarce and a mechanical connection had to be developed for the lamp. Since aluminum forms a thin oxide, this mechanical contact was not suitable for use on low-voltage lamps, such as flashlights and photoflash. The mechanical side connection required close tolerances between the glass seal and the metal base, which
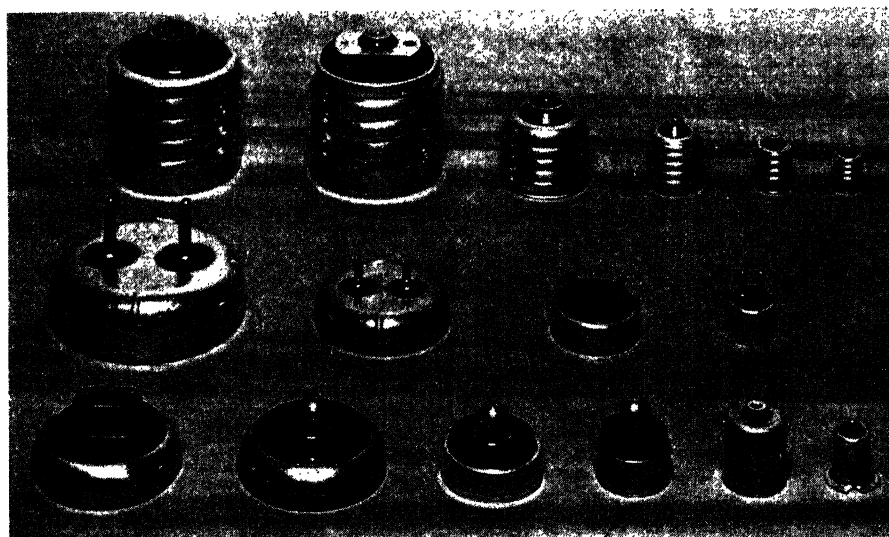
Fig. 1. Assorted incandescent and fluorescent lamp bases utilizing aluminum

Fig. 2. The voltage drop between the shell of the lamp base and the shell of the socket after exposure in corrosive atmosphere

proved to be very difficult to maintain in production. Recently, welders were developed and installed which now weld the side wire to the base. So much for some of the manufacturing and mechanical problems.

How does the new material perform in the field? One of the first questions asked concerning the aluminum base is, how does the oxide film affect the electrical connection? There is no single answer to this. On standard voltage lamps in a normal atmosphere, no adverse effect is found, probably because the voltage is high enough to break down the thin oxide film and the current required is very low compared to the area of contact. On low-voltage lamps, there are other factors. The voltage here is not sufficiently high to break down the thin oxide insulation. Many tests were run in humidity boxes, salt sprays, and outdoor atmospheres. Some were run on automotive vehicles with the lamps mounted under the bumper of a car, exposed to the salt and slush of Cleveland's streets. The conclusions are that the voltage drop between the shell of the socket and the shell of the lamp is practically zero when the lamp is first inserted into the socket. In a normal atmosphere, this will not change appreciably, but after the lamp has been exposed to adverse weather conditions for long periods, a voltage drop will develop but the magnitude is low, usually less than 0.2 volt. The voltage drop could be serious on low-voltage systems as in automotive service, where it would reduce light output and impair safety.

Fig. 2 shows the way the voltage drop develops with a bayonet base in an auto-

motive socket; also shown are the voltage drop for the treated staked-pin base, the extruded-pin base, and the lubricated extruded-pin base. In this type of test, lubricated extruded-pin aluminum bases compare very favorably with the present brass base. But more will be said about lubrication later. These data show how much can be done toward solving the basic service problem which aluminum presents.

There are other problems. Aluminum is considered to have a high coefficient of friction, primarily because the thin oxide film formed on the surface of the metal can be broken so easily. After the glassing-in process, some of the early aluminum bases were quite soft and could be deformed in subsequent operations. Many of these operations had to be readjusted to apply less pressure on the base. Some of the threads of these early bases were deformed by excessive pressure. A deformed base and the high coefficient of friction of the aluminum made these lamps difficult to insert in a standard socket.

The same condition exists where the tolerances between the base and the socket are very close. An abrading action takes place, breaking off some of the aluminum oxide, which in turn forms a good lapping compound. This further abrades the base. The same action occurs on a grip-type or antibackout type of socket where the lamp is purposely held securely by means of springs to prevent vibration from loosening the lamps in service. Fortunately, not many deformed bases or out-of-limit bases were marketed, and since there are not many of the grip-type sockets in use, the number of complaints received from the field was very low. The Company was still not satisfied with the "feel" of the lamp as it was inserted into the socket, and a lubricant was sought which would alleviate this condition. Commercial consideration demands that this material must be clean, odorless, colorless, and that it be a hard wax in preference to an oil or grease. The lubricant must not discolor at lampmaking temperatures
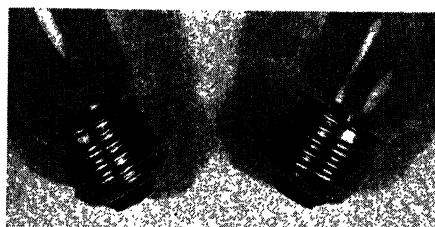
Fig. 3. The appearance of a treated aluminum base, at left, and a nontreated aluminum base

Magnesium and Alloys   CORRODED END
                                   (+)
Zinc
Aluminum
Cadmium

Duralumin

Steel, Iron
Cast Iron
Stainless (active)

Soft solders
Lead, Tin

Nickel, Inconel (active)

Copper
Brasses, Bronzes
Monel
Cupro-nickels

Nickel, Inconel (passive)

Stainless (passive)

Silver
PROTECTED END
(−)   Gold
                               Platinum

Fig. 4. A galvanic series of common metals

and, of course, should not vaporize at normal base-operating temperatures. It had to be inexpensive, and the method of application had to be such that it would not swallow the cost advantage that aluminum gave. The material that best met these specifications is the one now used by the Company, a hydrogenated whale oil. A very simple device was developed to apply the lubricant onto the base, and it does a satisfactory job as far as lubrication is concerned. Fig. 3 shows that the addition of the lubricant has not changed the physical appearance of the base, but it does improve its performance.

Another basic problem of aluminum is shown in Fig. 4, a galvanic series for some common metals and alloys. The metals grouped close together have little galvanic action, and the galvanic action increases as the two metals in question are separated in this series. Aluminum stands quite high in the galvanic series compared to brass. This in itself would be no detriment except that so many of the lamp sockets now in service have been manufactured with brass or copper shells. Aluminum-based lamps are used in these sockets. Aluminum in contact with the brass, and with water made conductive by atmosphere impurities, has a galvanic action. The aluminum is wasted away in this action and forms a deposit of aluminum oxide in the socket. If the material

A – UNTREATED AL. LAMPS – AL. SOCKETS
B – TREATED AL. LAMPS – BRASS SOCKETS
C – UNTREATED AL. LAMPS – BRASS SOCKETS
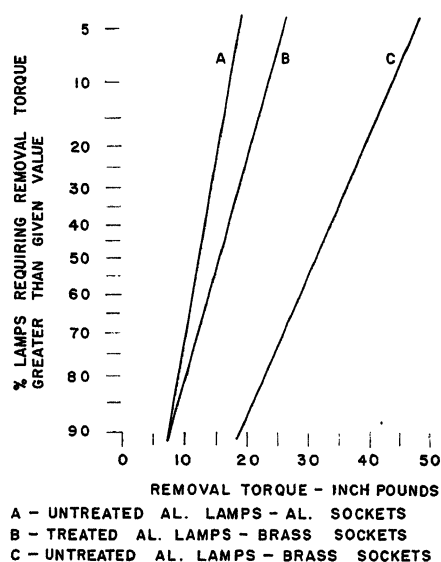
Fig. 5. Breakaway torque required to loosen lamps in outdoor-sign exposure

builds up to a sufficient volume, the lamp may be difficult to remove. This is particularly true in corrosive atmospheres of large industrial cities or in coastal areas when lamps are burned in such a way as to permit water to enter the socket.

If the lubricant used for reducing friction would now prevent or inhibit galvanic action, this special problem would be solved. That property was considered in the search for a suitable lubricant. Fig. 5 shows the test results on outdoor sign lamps, aluminum bases in copper or brass sockets that have been exposed to the elements for almost a year. The data show the breakaway torque required to remove lamps from the sockets. This is plotted for aluminum-based lamps in both brass and aluminum sockets and the lubricated aluminum-base lamps in brass sockets. As will be seen, the lubricant has inhibited corrosion to a large degree. Presumably there has been no galvanic action between the aluminum base of the lamp and the aluminum socket. The breakaway torque for the lubricated aluminum-based lamps in brass socket is not far removed from that for aluminum lamps to aluminum sockets, indicating that a good application of the lubricant will go a long way towards reducing or inhibiting galvanic action.

Fig. 6 shows the physical appearance of typical bases from this test. The lamps have been removed after almost a year's exposure in the sign test. The untreated base shows considerable corrosion, while the treated base shows very slight corrosion. In this galvanic action, the aluminum is the material which is

wasted away. The socket or the permanent part of the fixture is protected. The lamp is the expendable item; it has served its purpose and can now be discarded. Naturally, the Company is still on the lookout for materials which will afford better protection and better lubrication. Well over a billion lamps with aluminum bases have been made and on this production the number of complaints received is amazingly small. The complaints fall into two distinct types: those services having a humid atmosphere, where galvanic corrosion is a tremendous factor, and those where the dimensions or the mechanical features of the base and socket were such as to cause destruction of the oxide film on the aluminum with subsequent abrading and galling. The Company is satisfied that its lubricant will take care of the latter type, and is inclined to believe that the effectiveness of the lubricant in retarding galvanic action will vary with service conditions. It is realized that in certain applications the atmospheric conditions are unusually severe the lubricant is not adequate, and that brass bases should be used.

Better lubricants are constantly sought, not only for the present lamps but also to expand the use of the aluminum base to lamps of higher wattage. Household lamps are designed for use in paper-lined brass sockets and of necessity are designed for a low base-and-socket temperature. The Company's lubricant will be satisfactory under these conditions. However, there are lamps in higher wattages which have a much higher base-and-socket temperature and which are designed for use in porcelain sockets. In these cases, the lubricant in service at present would evaporate and leave a nonlubricated lamp in the socket. This will not cause trouble unless the socket is a friction- or grip-type device and/or the fit between the socket and the lamp is very close, in which case an abrading action will take place and the lamp might be difficult to remove. For these applications, a lubricant is needed which satisfactorily withstands these higher operating temperatures.

From a design point of view it was necessary to strengthen the aluminum base by using thicker materials than are used with brass, and this has been a source of trouble for some miniature lamps. Since the outside dimensions of the base must be kept constant, the increased metal thickness has reduced the inside diameter in the base. The reduced inside diameter has reduced the clearance which lead wires had, with the result that a satisfactory design for production on some of the miniature types has not yet been developed.

Another interesting side line is the single-contact bayonet base commonly used for the Company's automotive types and also on the midget photoflash lamps. Since this base was designed, many years ago, a straight wire pin staked into the brass shell has been used. Naturally, with the aluminum base exactly the same techniques were used. Trouble came from the fact that one flash gun holder for photoflash lamps had a spring clip which held the lamp by the pin of that base and made the electrical contact with this clip. The complaint was that lamps would not flash in the socket. An examination of the base revealed that the pin was completely insulated from the rest of the base. The low voltage of the flash gun was not sufficient to break down the thin oxide insulation, and the lamp was never energized. An examination of the product showed that a good percentage of the pins were so insulated. Only the thin layer of aluminum oxide was necessary to insulate this pin. An examination of brass bases showed that a large percentage of the pins had a very high resistance, on the order of 30 to 40 ohms between the pin and the shell. The answer to this was, of course, the adoption of the extruded-pin base which is now used. The metal of the shell is extruded to form the pin, and the result is an aluminum base which is better than the former brass base and which costs less.

Another example of an improved base by the use of aluminum is the miniature prefocused flashlight base, which was completely redesigned to

Fig. 6. The appearance of representative bases after outdoor exposure

A—Brass base
B—Nonlubricated aluminum base
C—Lubricated aluminum base
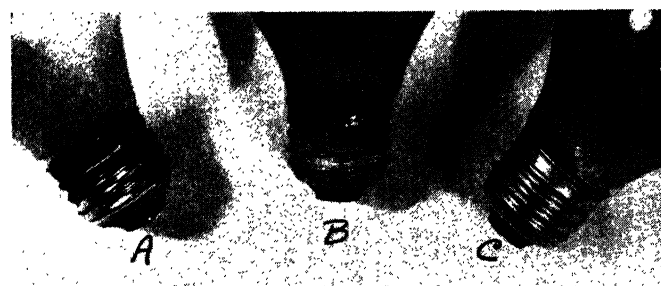
increase its strength, not merely by increasing metal thickness but also by changing some of the design features in the base. The result today is that under the Company's form of testing, the prefocus flashlight base is stronger in aluminum than it was in brass. The end result is a lamp which gives satisfactory service to the customer at an over-all saving in lamp manufacturing cost. The net saving per lamp is, of course, a very small fraction of a cent and is not felt by the customer in the form of lower prices. But, with other savings, it tends to offset the higher labor cost that has been prevalent in the past 10 years and it has made it possible to defer price increase for lamps for a time.

The long-term objective of the Company is to expand the use of aluminum in the lamp industry as fast as engineering skill produces solutions to the problems in terms of consumer applications.

Great progress has been made but there is room for more, both in redesigning the product to perform satisfactorily, and in educating the consumer to accept the new base. The aluminum base is serving well and it should no longer be stigmatized by being called a substitute. It has limitations but some of them have already been overcome. Further developmental effort will result in increased uses for aluminum in the lamp industry.

Design and Protection of A-C Power Systems for Aircraft Instruments

OSCAR MARKOWITZ
MEMBER AIEE

CHARLES SEARS
ASSOCIATE MEMBER AIEE

THE role of modern military aircraft is inexorably reflected in its electric system. The aircraft electric system must maintain its continuity to complete its mission. Beyond that, the utmost of reliability and continuity is required of the electric power system supplying "safety of flight" items such as the Gyro horizon instrument, the turn and bank instrument, the fuel quantity instrument, the engine instruments, and the fuel flow instruments. These instruments when inoperative can easily cause loss of an aircraft.[1]

In the past few years much concentration has been put into the design of electric systems for flight instruments, bringing to light peculiarities inherent to the small low power-factor loads supplied through high-impedance, low-capacity power sources. In this paper many of the problems and design aspects are defined which resulted from efforts made at the U. S. Naval Air Development Center (NADEVCEN) to increase the reliability and protection of flight instrument power systems.

Design Problems

The most vital of electrically operated instruments in aircraft are usually duplicated in one form or another. For utmost in reliability, power supplied to each duplicate instrument is made as independent as possible back to a source. The thought of power system independence for functionally duplicate instruments usually requires compromise of various degrees with different aircraft configurations. In addition, to increase reliability of the instrument power system, power is selected from more than one source. This compromise and/or selection between multisources of power then imposes the need for switching in these circuits from one source of power to an alternate source and, in turn, the requirement to sense the need for such transfer.

It is extremely important that the pilot be aware immediately of any instrument failure and the need for power transfer. Efforts are being made to provide indication of power and/or instrument failure inherent with the instrument.

However, most dependence is placed on relays to detect failure in the power system and in turn provide a warning signal or initiate an automatic transfer to another power source.

Start of design for instrument power is made from a load analysis wherein the power requirements are set and the power sources are determined. Instruments require small amounts of power, e.g., 5 to 10 volt-amperes (va) at a low power factor for the a-c gyro-operated instruments. Laboratory experience has shown that basic manufacturer's information on instrument power requirements has not been adequate. Some instruments are 3-phase balanced or unbalanced loads, others single-phase loads, making complete data most important for properly phasing the loads to present, insofar as possible, a total balanced 3-phase requirement. Measurement of aircraft flight instrument power requirements is complicated by the effects of meters upon circuit conditions being measured when using 1/4-per-cent meters for 400 cycles. Lack of realization of the importance of accurate load data and absence of military requirements for the same have resulted in inaccuracies which prevented system designers in achieving desired reliability goals for instrument systems.

Instruments are usually supplied from a 3-phase delta, 115-volt 400-cycle power with one line grounded. Sources of power vary from 100-va inverters and small direct driven a-c generators to larger inverters and primary 115/200-volt wye power systems with appropriate transformers. Most instrument power systems contain both the heavier power system for normal operation and the smaller inverters as alternate power for the vital instruments during conditions of limited power and/or emergency operation. Field experience has painfully demonstrated that, when more than one source of power is available, both sources must be operated continuously. Idle inverters during flight operations have proved to be less reliable than the same inverter used continuously. The emergency which requires the use of the alternate power source should occur very seldom, perhaps once during the life of the

aircraft. When the emergency does occur, however, it should be known that the alternate power source will perform properly. One way of being sure is to have it perform a continuous function with some useful load. Thus, failure of the alternate source is apparent at all times and is not dependent upon emergency operation for detection.

Facility in changing from one source of power to an alternate source has been considered essential for increased reliability of the vital instruments. In the past 5 years, design trends have changed from manual switching to automatic and now back to manual. Automatic switching failed to provide the required reliability since the fault-sensing relays were not sufficiently developed to detect all fault conditions, while experience showed transfer relays to be less reliable than the rest of the system. This has resulted in the present design to reduce the components in these most vital power systems to the barest minimum required for manual power and/or instrument transfer. System design in the past has not placed sufficient emphasis on pilot requirements leading toward a goal of only one simple control manipulation during an emergency to accomplish all required disconnections and transfers. This can provide critically needed lead time while the pilot must decide the nature of his emergency and initiate corresponding corrective actions.

Problems of fault-sensing and protection are aggravated by switching from one power source to another of different capacity. Large a-c systems and the larger inverters have regulation sensed from an average of the 3-phase voltages. All 250-va inverters have regulation sensed from a single phase; 100-va inverters have no inherent voltage regulator. This leads to complication in making any given failure-sensing and fault protection compatible to different sources of power, with such a wide range of source impedance and regulation characteristics.

Load Analysis

Present method for load analysis according to Specification No. MIL-E-7016[2] utilizes basic 3-phase va values of which volts, amperes, and watts are measured

Table I. Sample Load Analysis of Instrument Power Requirements in Accordance With MIL-E-7016

A EQUIPMENT	B PART DESIGNATION	C TOTAL NO. OF UNITS	D OPERATING TIME IN MIN	E ELECTRICAL REQUIREMENTS PER UNIT 115 V 3 PHASE — E1 VA	E2 WATTS	E3 VARS	26 V 1 PHASE — E4 VA	E5 WATTS	E6 VARS	F PF	G VOLT REG	H FREQ RANGE	I CONNECTED LOAD — I1 WATTS	I2 VARS
PILOT'S STANDBY														
G-2 COMPASS		1	30	48.2	36.5	31.7				0.76		380-420	36.5	31.7
FLIGHT INSTRUMENTS														
GYRO HORIZON IND.	R88-1-1325-11	1		5.5	3.8	3.9				0.69			3.8	3.9
TURN BANK IND.	R88-1-3150	1		4.3	2.7	3.3				0.63			2.7	3.3
FAILURE IND.	PR-36200	1		3.2	2.0	3.0				0.62			2.0	3.0
TRANSFORMER (LOSSES)	R17-T-7218-235	1		4.5	4.0	2.1				0.89			4.0	2.1
ENGINE INSTRUMENTS														
DUAL FLOWMETER IND.	R88-1-1200-12	1					2.60	0.45	2.55	0.17			0.45	2.55
FLOWMETER TRANSM.	R88-T-1965	2					5.18	0.90	5.10	0.17			0.90	5.10
DUAL MANIF. PRES. IND.	R88-1-1585-12	1					2.34	0.54	2.27	0.23			0.54	2.27
MANIF. PRES. TRANSM.	R88-T-2350	2					2.34	0.54	2.27	0.23			0.54	2.27
DUAL JET OIL PRES. IND.	R88-1-1925	2					4.74	1.08	4.54	0.23			1.08	4.54
JET OIL PRES. TRANSM.	R88-T-2651-100	4					4.74	1.08	4.54	0.23			1.08	4.54
REPEAT IND. CIRCUIT														
MASTER DIR. IND.	R88-1-1681-25	1					4.92	3.65	3.30	0.67			3.65	3.30
NAVIG. REPEAT IND.	R88-1-803	1					2.06	1.53	1.38	0.67			1.53	1.38
TRANSFORMER (LOSSES)	GE 68G177	1	30	7.0	4.0	5.9				0.57		380-420	4.00	5.90
TOTAL				72.2	53.0	49.9	27.60	9.77	25.95				62.8	75.9

VA = 98.5

NOTES:
⚠1 DATA FROM MANUALS
⚠2 ESTIMATED REQUIREMENTS
⚠3 26 V 3φ INSTRUMENTS
⚠ ESTIMATED 3φ DISTRIBUTED TRANSFORMER LOSSES

with reactive power and the power factor calculated. This method lends itself to simplified analysis for balanced wye configurations with grounded neutral. However, when combinations of unbalanced loads, transformations, and delta configurations are involved, the vector line currents govern calculations and analysis. With instrument power systems, within the frame of the *MIL-E-7016* specification, all operations become cumbersome and inaccurate, while identity and degree of unbalance in an individual line are lost. Thus, it is difficult to determine possible single-phase overload of a small inverter when total va are within the machine rating.

Table I illustrates a load analysis made in accordance with the *MIL-E-7016* specification while Table II illustrates another approach in analyzing the same loads. For Table II, all values were measured as currents and voltages with their corresponding angles. These values were converted into corresponding 3-phase complex notation. Sums of the individual entries were used to determine corresponding magnitudes of line currents. It is readily apparent that the increase of facility and accuracy of the method illustrated in Table II revealed the overload on one phase of the inverter. While full load per phase at 100 va is a 0.5-ampere line current, the load analysis in Table II showed line currents of 0.33, 0.62, and 0.77 ampere respectively. The I_c line current of 0.77 ampere represents a 50-per-cent overload on the corresponding wye winding of the *E-1616* inverter, and was not suspected from Table I and its calculated full load of 98.5 va.

During a load analysis with a contemplated instrument power system design, it became apparent that a 3-to-1 phase transformer created more unbalance than a corresponding single-phase transformer operating from one of the three phases. Fig. 1 shows phasor diagrams of these load conditions. This would have been completely missed without the use of complex notation for current in the load analysis. Use of the single-phase transformer permitted the selection of phase connection that provided the best balance. Other experiences have indicated that with any unbalance extreme care is required when using a 250-va inverter with voltage regulation from single-phase sensing. In Fig. 1(A) the line voltage V_{AB} has become excessive because of the unbalance. Phase loads must be assigned judiciously to minimize the 1-phase regulation effects. In any case with unbalanced loads, it is strongly recommended

Table II. Sample Load Analysis of Instruments Shown in Table I Using Measured Values and Complex Notation

A EQUIPMENT	B PART DESIGNATION	C TOTAL NO. OF UNITS	D OPERATING TIME IN MIN	E VOLT REQ'D	F FREQ. RANGE (CPS)	I_A	I_A	I_B	I_B	I_C	I_C	H_1 (VA)	H_2 (W)	H_3 (VARS)	H_4 (VA)	H_5 (W)	H_6 (VARS)	I PF	J_1 (VA)	J_2 (W)	J_3 (VARS)
PILOT'S STANDBY																					
G-2 COMPASS			30	115±5	380–420	0.11	-0.10-J0.06	0.33	+0.02-J0.33	0.39	+0.12-J0.37	50.8	45.1	23.3				0.89		45.1	23.3
FLIGHT INSTRUMENTS																					
GYRO HORIZON IND.	R88-1-1325-11	1		26		0.07	-0.06+J0.03	0.07	+0.04+J0.05	0.07	+0.02-J0.07	5.5	3.8	3.9				0.69		3.8	3.9
TURN BANK IND.	R88-1-3150	1		26								4.3	3.7	3.3				0.63		3.7	3.3
FAILURE IND.	PR-36200	1		26								3.2	2.0	3.0				0.62		2.0	3.0
TRANSFORMER (LOSSES)	R17-T-7218-235	1		115/26								4.5	4.0	2.0				0.89		4.0	2.0
ENGINE INSTRUMENTS																					
DUAL FLOWMETER IND.	R88-1-1200-12	1		26											4.11	1.54	3.80	0.37		1.5	3.8
FLOWMETER TRANSM.	R88-T-1965	2		26																	
DUAL MANIF. PRES. IND.	R88-1-1585-12	1		26											5.56	1.81	5.26	0.32		1.8	5.3
MANIF. PRES. TRANSM.	R88-T-2350	2		26																	
DUAL JET OIL PRES. IND.	R88-T-1925	2		26		0.38	-0.26-J0.27	0.47	-0.14+J0.45	0.52	+0.09-J0.51				18.4	5.22	17.64	0.28		5.2	17.6
JET OIL PRES. TRANSM.	R88-T-2651-100	4		26																	
REPEAT IND. CIRCUIT																					
MASTER DIR. IND.	R88-1-1681-25	1		26											3.9	3.12	2.37	0.79		3.1	2.4
NAVIG. REPEAT IND.	R88-1-803	1		26																6.6	2.7
TRANSFORMER (LOSSES)	GE-680177	1	30	115/26	380–420							7.1	6.6	2.7				0.93			
TOTAL						0.33	-0.23-J0.24	0.62	+0.02+J0.62	0.77	+0.09-J0.76	75.6	65.2	38.2	31.3	11.69	29.07	0.93	101.3 / 106.1	76.9 / 85.7	67.3 / 62.2

⚠ TOTAL OF INDIVIDUAL LOADS

⚠⚠ MEASURED TOTAL (LESS FAILURE INDICATOR) WITH E-1617 INVERTER

TOTAL ESTIMATED POWER REQUIREMENT

⚠ CALCULATED FROM ⚠ VOLTAGE REFERENCE V_{CA} (124+J0)

that a laboratory mockup be required to provide final answers as to the best distribution configuration.

There is always a strong temptation to correct the low power factor of instrument loads with capacitors. However, capacitors for power correction cannot be applied indiscriminately to instrument power systems. This is caused by the high source and distribution impedances which, when combined with the capacitors, readily form resonant circuits at frequencies in the order of third and fifth harmonics of the 400-cycle power. For example, capacitive correction used in the P-1 autopilot was checked to determine its effect on harmonic current in the system by varying the capacity and analyzing the line currents for harmonic content. Fig. 2 shows the results obtained. The actual resonant point for the fifth harmonic was approximately 3.2 microfarads while the value of the P-1 corrective capacity was 4 microfarads. At 3 microfarads, the fifth harmonic current was 0.62 ampere compared to 1.0 ampere of fundamental. Without any corrective capacity, the fifth harmonic was 0.027 ampere compared to 2.075 amperes of fundamental.

Power Sources

Paragraph 3.1 in Specification No. MIL-E-7894[3] defines the requirement for multisources of power with duplicating functions. This is carried further for the safety-of-flight items to result in multisources of power for those items. The requirement reflects itself into instrument power systems as two different capacity sources when two different instrument power requirements exist for normal conditions and alternate (emergency) conditions. Thus, for normal conditions, the power source can be a 250-va or higher capacity inverter or the primary a-c source of 9 kva or higher, whereas the alternate source is usually a 250-va or 100-va inverter. Because of using more than one power source with different characteristics, special problems are created peculiar to instrument power system design.

Fig. 3 shows part of an instrument power system which could be powered from 250-va (Navy-type E-1617) or 100-va (Navy type-E-1616) inverters. Short circuits in the 26-volt a-c system, using the E-1616, caused a drooping inverter output characteristic. However, for the E-1617, the same short-circuit conditions caused line voltage rises. The E-1617 has single-phase voltage regulation and when the regulated lines are required to supply overload currents, the voltages
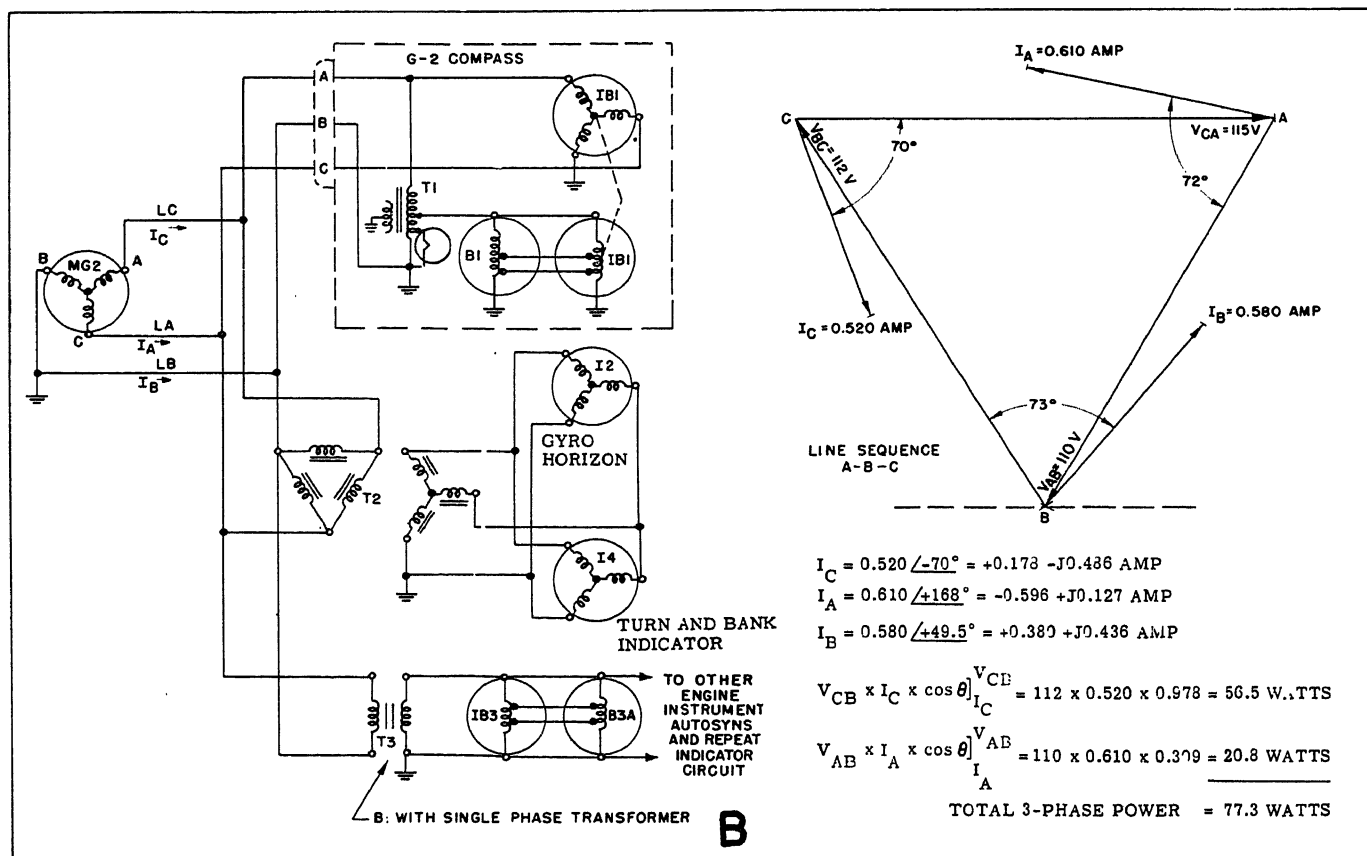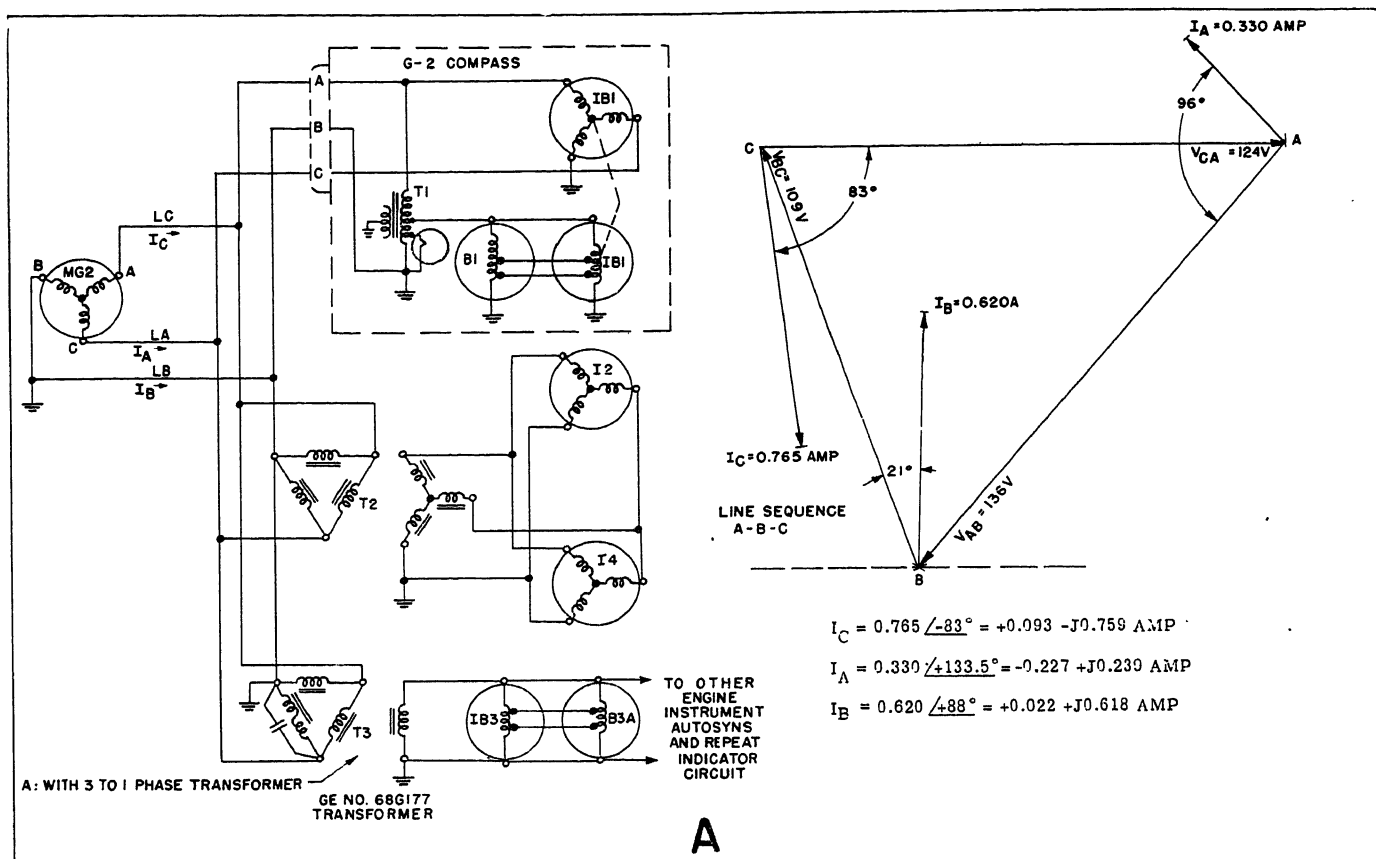
Fig. 1. Improvement in balance when 3-to-1 transformer is replaced with a single-phase transformer

from those lines to the third line rise to excessive values. For example, with the 250-va inverter one set of conditions recorded during a short circuit were:

$V_{AC} = 186$ volts, $V_{58} = 33.0$ volts
$V_{CGN} = 174$ volts, $V_{76} = 31.5$ volts
$V_{AGN} = 114$ volts, $V_{87} = 0$ volt

The changes in primary to secondary voltage ratios are caused by the excessive internal transformer losses with a redistribution of phase loading. Yet, with the 1,500-va (Navy-type *E-1737*) inverter, the same fault condition did not load the inverter to the extent that its output voltages were affected materially.

During selection of inverters, consideration should be given to its short-circuit capacity. Reference 4 provides details of tests made on various inverters. The tests revealed that some inverters have short-circuit capacities less than their full load current under short-circuit conditions.

Fault Protection and Co-ordination

Reference 5 describes a mockup of the instrument power and distribution system shown in Fig. 4 for tests under fault conditions. The specific faults investigated in this system were:

1. Open secondary lines.

2. Open primary lines.

3. Uninterrupted short-circuited secondary, line-to-line.

4. Uninterrupted short-circuited secondary, line-to-ground.

5. Uninterrupted short-circuited secondary, symmetrical 3-phase.

6. Interrupted short-circuited secondary, line-to-line.

7. Interrupted short-circuited secondary, line-to-ground.

8. Interrupted primary line for a line-to-line secondary short circuit.

9. Interrupted primary line for a line-to-ground secondary short circuit.

10. Interrupted primary line for a symmetrical 3-phase secondary short circuit.

The faults were applied at various locations in the flight instrument circuits. In general, the faults applied were open and short circuits that involved the different lines of the 3-phase system. One set of data is shown in Fig. 5 of the CPGHID circuit blocked in Fig. 4(A). The short circuits tested were subdivided into the interrupted and the uninterrupted faults. The interrupted short-circuit tests were made to study those conditions where only one line fuse would open for a particular short-circuited section. The uninterrupted short circuits
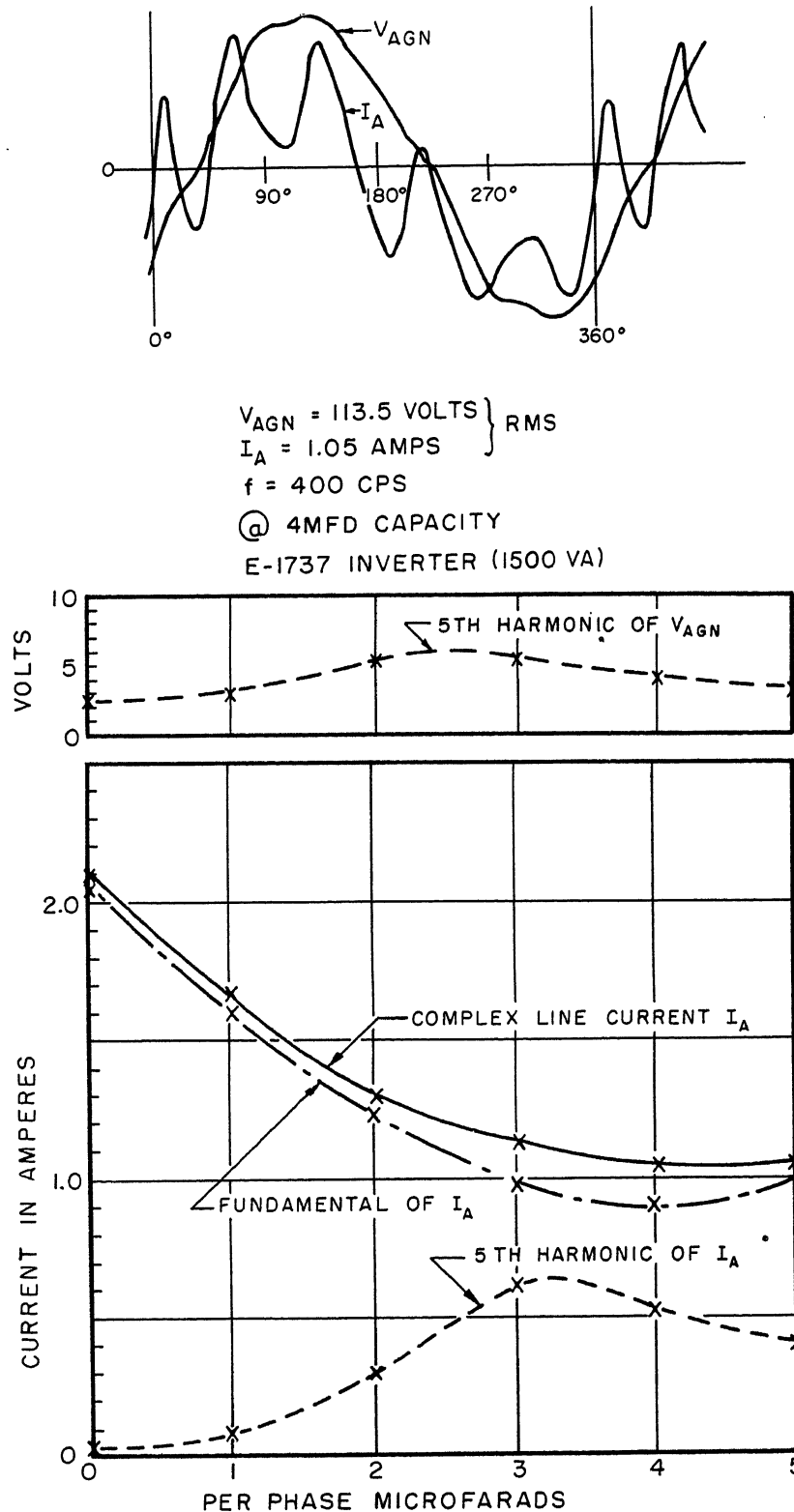
Fig. 2. Effect of corrective capacity in P-1 autopilot on fifth-harmonic content of line current and line voltage

were made to study the effects in the system while simulating no fuse or no interruption by a fuse.

Fig. 6 shows the correlation between the comprehensive data of tests performed on the system shown in Fig. 4 and the fuses selected. The following distinctive problems were encountered in selecting these fuse values:

The starting currents for certain instruments were equal to or greater than the currents with fault conditions present. Such a condition is shown at A in Fig. 6. A fuse selected to protect against this fault would also interrupt during a normal start. In addition, consideration must
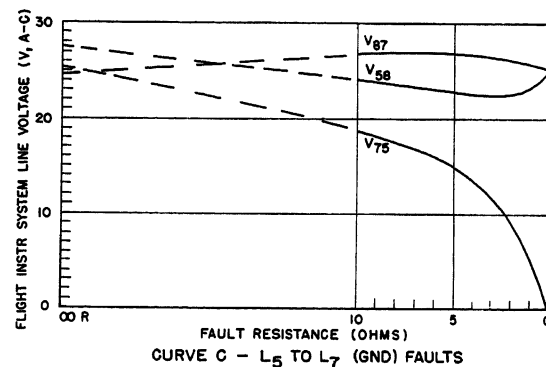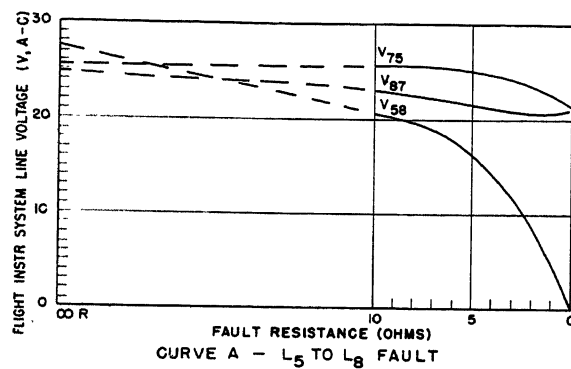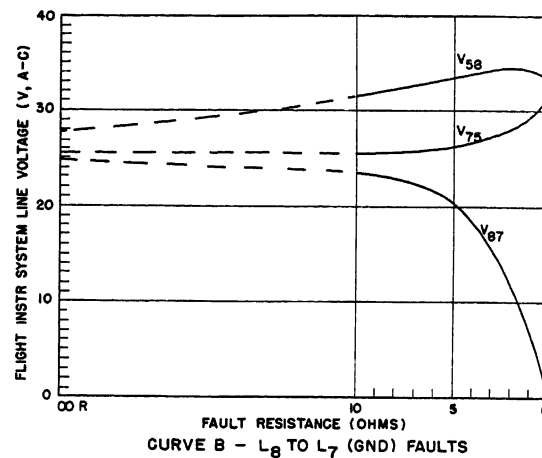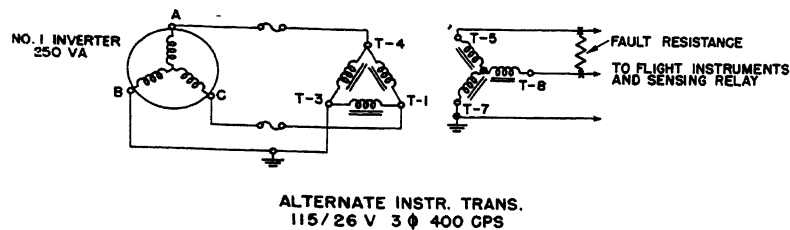
ALTERNATE INSTR. TRANS.
115/26 V 3φ 400 CPS

CURVE B — L_8 TO L_7 (GND) FAULTS

CURVE A — L_5 TO L_8 FAULT

CURVE C — L_5 TO L_7 (GND) FAULTS

Response to Short Circuit Faults Using No. 1 Inverter (250 VA)

ALTERNATE INSTR. TRANS.
115/26V 3φ 400CPS

CURVE B-L_8 TO L_7 (GND) FAULT

CURVE A-L_5 TO L_8 FAULT

CURVE C-L_5 TO L_7 (GND) FAULT
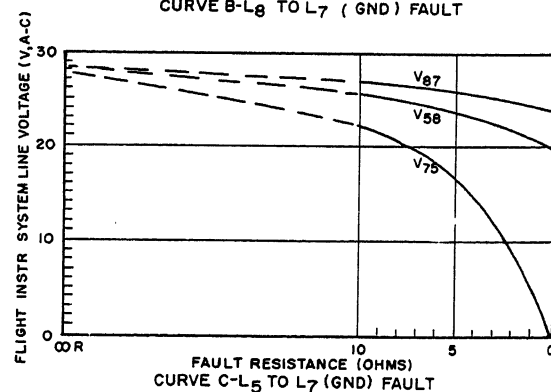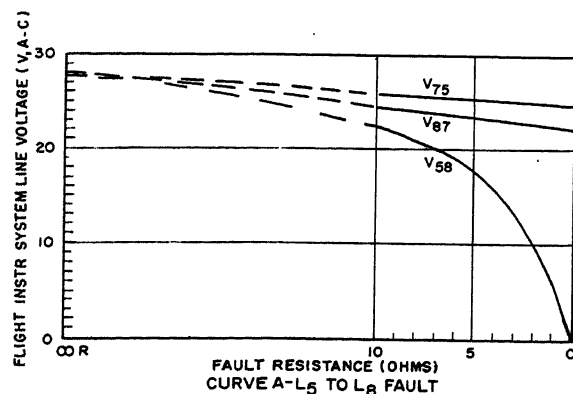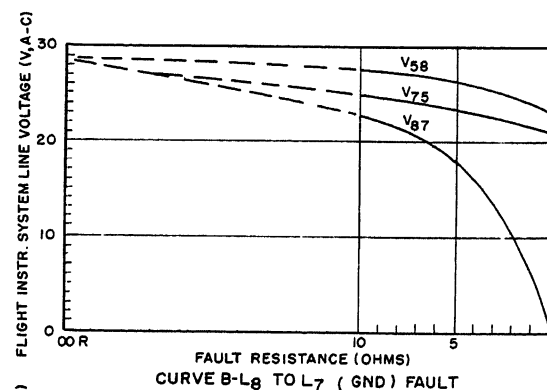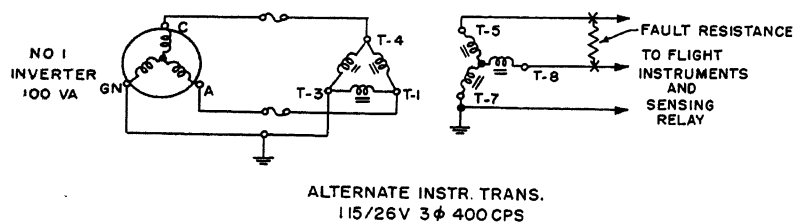
Response to Short Circuit Faults Using No. 1 Inverter (100 VA)

Fig. 3.   Effect upon 3-phase voltages with different inverters and with a given type of fault

be included for the more severe start current-time characteristics of gyro instruments at low temperatures. Reference 6 provides data for such start conditions. A sample of these data is abstracted and shown in Fig. 7.

Fuse co-ordination is required between primary and secondary of the transformer circuits. When a fault occurs in the secondary circuits of the transformer, it is essential that only the faulted section be isolated to minimize the effect of the fault on other vital instruments. Co-ordination becomes a variable because of low and different source capacities, high transformer equivalent series impedance, redistribution of highly unbalanced 3-phase currents through the transformer windings, and much increased transformer losses during fault conditions. The changed spreads of line currents corresponding to the primary and secondary of a transformer with a line-to-ground fault in the secondary is shown as *B1* compared to *B2* in Fig. 6. Appendix *B* of reference 5 details the co-ordination relationships for the system in Fig. 4.

In the 3-phase circuits of the system, a fault on a line may open the fuse in that line but, in the second line, abnormal current will continue to feed the same fault by feedback through instruments. The magnitude of this current is higher than normal but does not exceed starting current values and cannot be protected for effectively. Such a condition is shown in *C* of Fig. 6. Although it is unlikely to cause fire and smoke, it will cause the instrument to malfunction sooner than if both lines had been opened by the fuses.

Selection of fuse values and types should be made carefully to assure mechanical reliability.[7]

For any given fault, the currents are different for different inverters. The *E-1616* inverter has no voltage regulator. The *E-1617* inverter has a voltage regulator sensed from one phase. A fault on the regulated phase will create overvoltage on the other phases. *D* of Fig. 6 shows different currents with different inverters and the same fault condition.

Fault Sensing

Protection serves its purpose in clearing a fault but, in addition, the pilot requires some warning that his instruments have become unreliable. The warning can be in the form of an individual instrument indicator such as a flag and/or a panel warning light operated from a sensing device. When voltage-operated 1-, 2-, or 3-phase power-failure relays are used for this purpose, sensing is not reliable

FGT	FLUX-GATE TRANSMITTER	SR	SENSING RELAY
PTBID	PILOT'S TURN & BANK INDICATOR	CPRDID	CO-PILOT'S REPEAT DIRECTION INDICATOR
PGHID	PILOT'S GYRO HORIZON INDICATOR		
PMDID	PILOT'S MASTER DIRECTION INDICATOR	CPGHID	CO-PILOT'S GYRO HORIZON INDICATOR
NRDID	NAVIGATOR'S REPEAT DIRECTION INDICATOR	K1	REMOTE-CONTROL RELAY (PILOT'S INSTR)
MI	MAIN INVERTER	K2	REMOTE-CONTROL RELAY (CO-PILOT'S INSTR)
SI	STANDBY INVERTER		

Fig. 4. Single-line diagram of a flight instrument a-c electric system for aircraft

unless power failure is almost complete. This method of fault sensing is generally inadequate since some faults, although depressing voltages sufficiently to make instrument operation unreliable, do not provide sufficient voltage differential for relay operation. In addition, with an open circuit, voltages do not collapse completely because of feedback through instruments or transformer.

Numerous fault tests were made in the instrument portion of the *P-1* autopilot system, Fig. 4(B), to determine fault-sensing adequacy of an integrated 2-phase voltage-sensing relay. These tests also included the effects of different inverters as power sources. The relay was a 26-volt unit connected as shown at *C* of Fig. 4. Cause, effects, and results of these tests are tabulated in Table III. Only 12 per cent of all fault conditions resulted in immediate good indication, while 46 per cent were delayed or unreliably indicated. A delayed or unreliable indication occurred when voltages in the system decreased gradually, as shown in Fig. 8, for an open secondary line, and for an open primary line. Further unreliable aspects occurred when a fault, after interrupting a line fuse, was removed within 30 seconds. The system voltage partially recovered and remained at a value sufficient for relay pickup, thus losing the initial indication. This aspect was representative of a typical service condition since aircraft faults are known to be of an intermittent nature and are not usually sustained. The re-
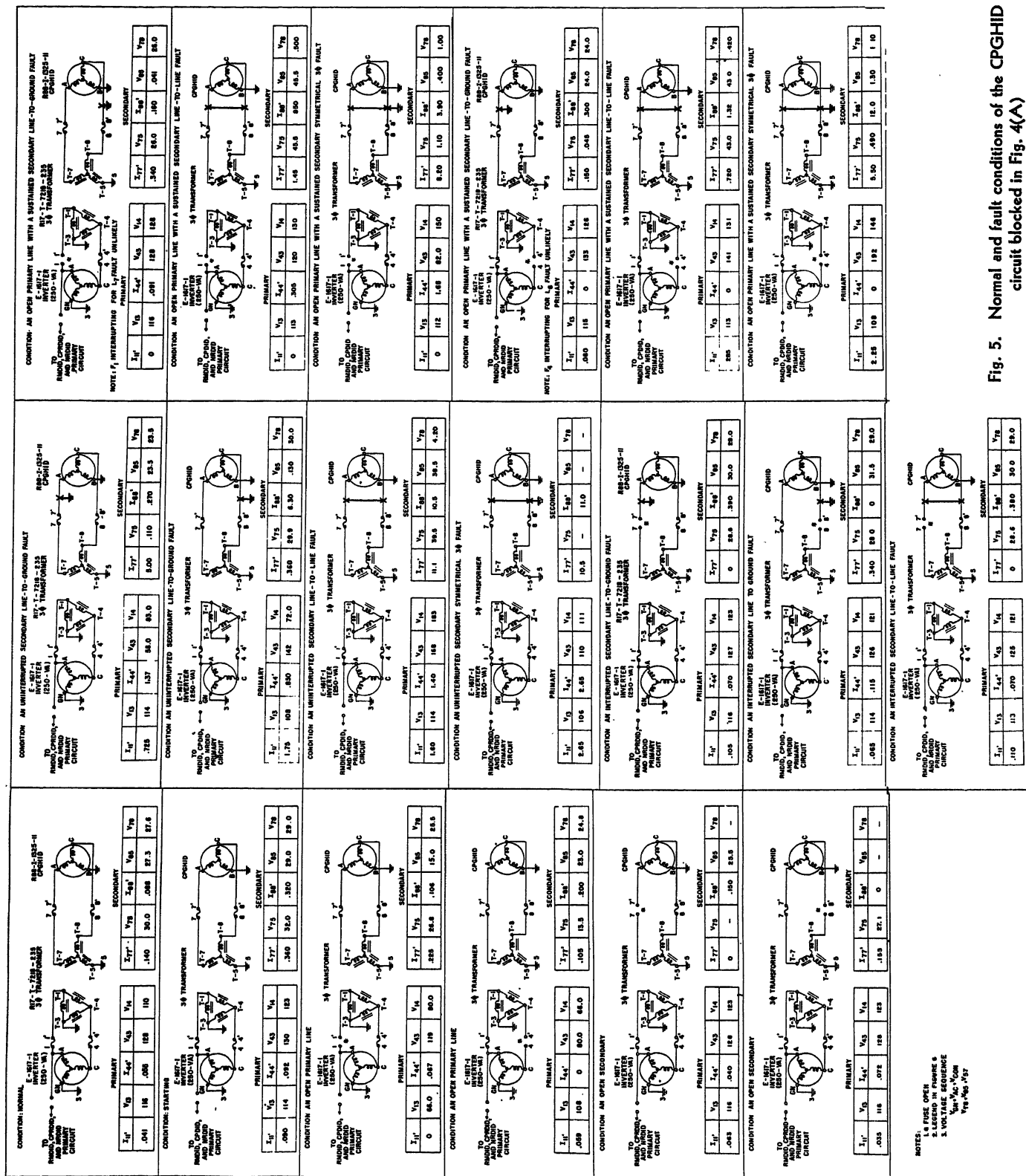
Fig. 5. Normal and fault conditions of the CPGHID circuit blocked in Fig. 4(A)

maining 42 per cent of all the faults definitely were not indicated. Reference 8, from which the foregoing data were obtained, details these tests with corresponding results and concludes that the integrated 2-phase sensing relay is inadequate as a fault sensing relay for aircraft instrument power systems.

Any relay which must operate upon such small voltage changes will require critical tolerances and will be unsuitable for fault sensing. Development of a relay with a different approach is required. Adequate fault sensing is a basic essential element for either manual or automatic transfer of power. Until the sensing can provide more intelligence than the pilot relative to the power condition and/or operating condition of the instruments, automatic transfer to alternate power is not consistent with the order of reliability required of the instrument power system.

Conclusions

In conclusion, it is felt that the following items should be considered as important during the design stages of an aircraft instrument power system:

1. Power systems for safety-of-flight in-
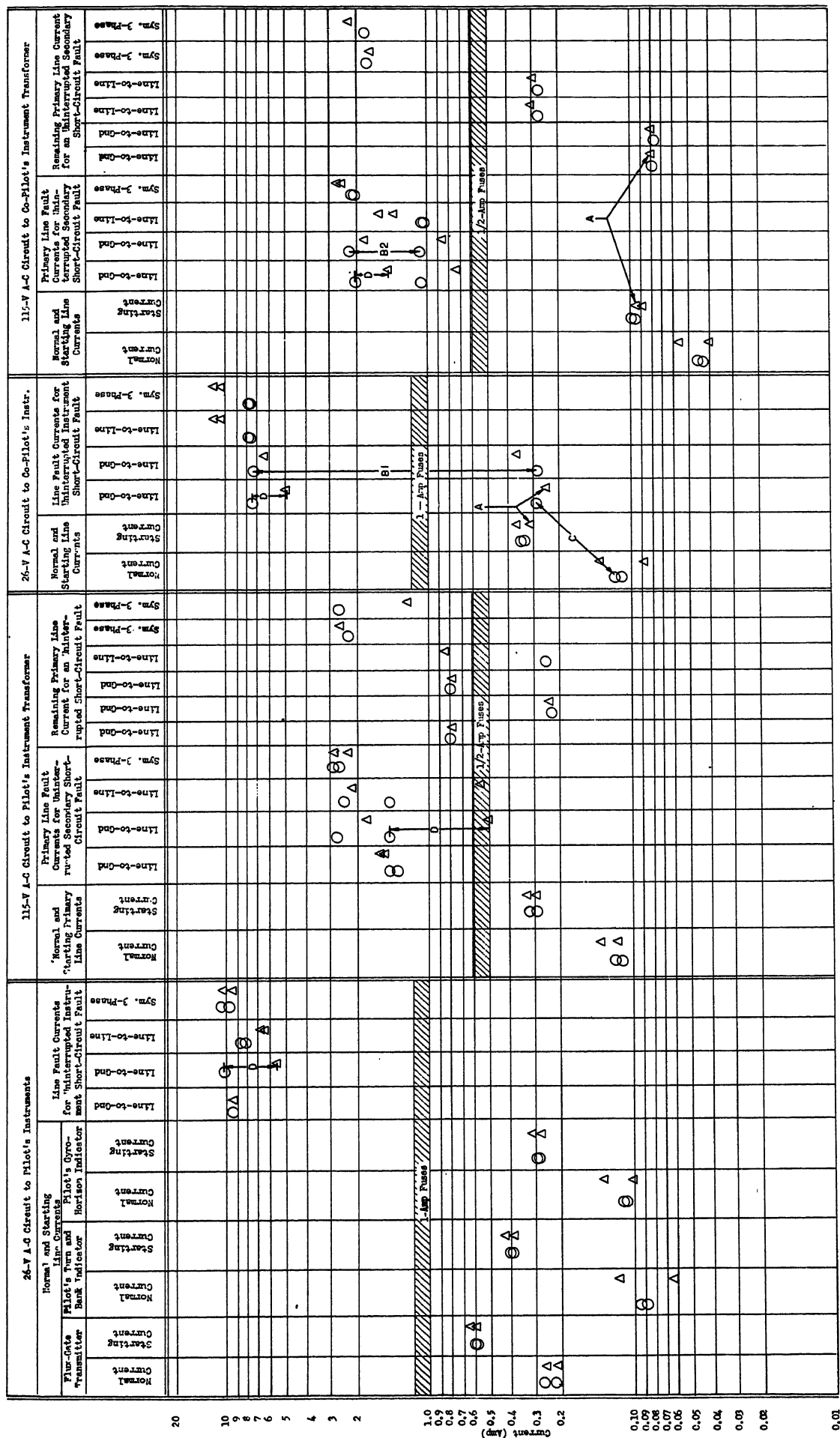
Fig. 6. Current magnitudes for various normal and fault conditions obtained in the system shown in Fig. 4 compared to fuse selections

O   Individual Line Current for Main Inverter Operation
△   Individual Line Current for Standby Inverter Operation
▨   Estimated Fuse Operation for 100 to 125 Percent of Rating
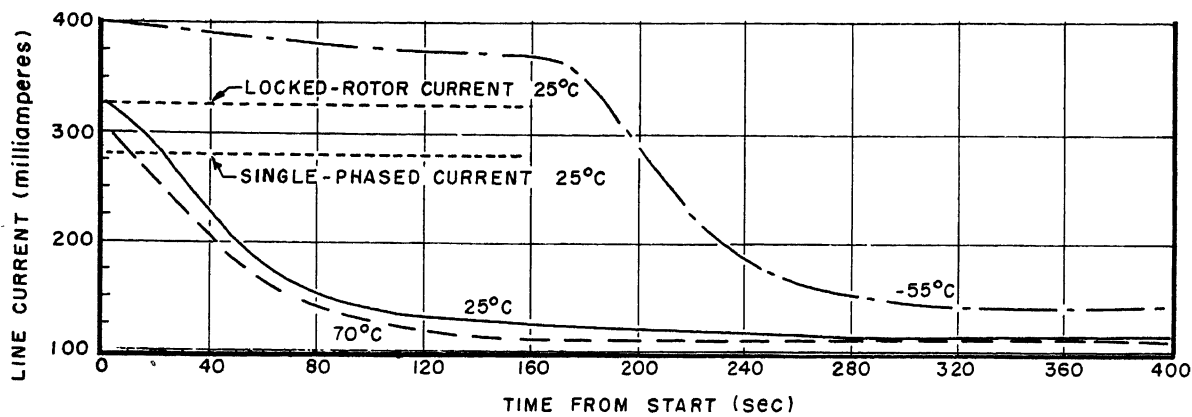
Fig. 7. Starting current characteristics of gyro horizon indicator

Figure axes: LINE CURRENT (milliamperes) vs TIME FROM START (sec)

Curve labels: LOCKED-ROTOR CURRENT 25°C, SINGLE-PHASED CURRENT 25°C, 25°C, 70°C, -55°C

struments should be designed to contain parallel sources of power as completely independent of each other in so far as practicable to supply corresponding functionally duplicated instruments. If possible, these power sources should provide instrument power independently of the main electrical system.

2. Idle and/or inoperative equipment

should be avoided as an alternate source of power.

3. The load analysis should be made in greater detail than called for by Specification No. MIL-E-7016 so that a complete and accurate analysis of line currents in the 115-volt delta system can be obtained, and every effort should be made to present a balanced 3-phase load.

4. Application of 3-phase to 1-phase transformers should be made only after load analysis to see if they are acceptable, considering 3-phase balance and power factor.

5. Where inverters are used as power sources, selection should be made to assure sufficient short-circuit capacity.

6. When the slow-blow type of fuses are applied to the design, they should be selected

Table III. Cause, Effect, and Results on P-1 Instruments With Response of a 2-Phase Relay During Fault Conditions in the Circuit Blocked as B of Fig. 4

Cause	Effect	Results	
Fault	Electrical Failure	Flight Instrument Response	Integrated 2 Φ Sensing Relay Response
Open line to GHID. Broken wire or connection. Broken fuse (mechanically). Interrupted fuse (transient overload).	Unbalanced voltages to GHID and sensing relay.	GHID tumbled in 4 to 5 minutes. Other instruments unaffected.	No indication, line 4 or line 6 involved. Indication, line 5 (GND) involved. 5- to 7-second delay.
Open line to all instruments. Broken wire or connection.	Unbalanced voltages to all instruments and sensing relay.	GHID tumbled in 3-1/4 minutes. FGT and MDID malfunction from 2 minutes until FGT gyro tumbled in 3-1/4 to 15 minutes.	No indication, line 4 or line 6 involved. Indication, line 5 (GND) involved. 10-second delay Response different for E-1617-1 inverter.
Interrupted short circuit at GHID (incomplete isolation of faulted section where one line fuse interrupts)	Unbalanced voltages to GHID and sensing relay.	GHID tumbled in 1-1/2 minutes. Other instruments unaffected.	No indication, line 4 to line 6 involved. Indication, line 6 to line 5 (GND) or line 4 to line 5 (GND)
Interrupted short circuit at GHID. Short circuit less than 30 seconds duration.	Unbalanced voltages to GHID and sensing relay. Change in unbalanced voltages when short circuit is removed.	GHID tumbled in 2 to 4 minutes. Other instruments unaffected.	During period of short; no indication - line 6 to line 4, indication - line 6 to line 5, indication - line 4 to line 5. When fault was removed within 30 seconds the indication was lost.
Uninterrupted short circuit as in unprotected circuits.	Unbalanced voltages to all instruments and sensing relay.	GHID tumbled in 2 minutes. FGT and MDID malfunction from 1 minute until FGT gyro tumbled in 1-3/4 to 3-1/2 minutes.	No indication, line 4 or line 6 involved. Indication, line 6 to line 5 (GND) or line 4 to line 5 (GND)
Uninterrupted short circuit where the resistance of the fault is more than 1-1/2 ohms*.	Unbalanced voltages to all instruments and sensing relay.	No instrument response data obtained as a function of fault resistance.	No indication.
Open line to transformer primary circuit. Broken wire or connection. Broken fuse (mechanically). Interrupted fuse (transient overload).	Unbalanced voltages to all instruments and sensing relay.	GHID tumbled in 3-1/2 to 14 minutes. FGT and MDID malfunction from 2 minutes until FGT gyro tumbled in 3-1/2 minutes.	No indication, line 1 involved. Indication, line 3 or line 2 (GND) involved, 18- to 20-second delay
Inverter failure lock rotor or loss of D-C supply to inverter	Loss of all voltages to the instruments and sensing relay.	GHID tumbled in 3 minutes. FGT tumbled in 4-1/2 minutes.	Indication.
Inverter regulator failure causing 3-phase reduction of output voltages.	Equal undervoltage of all three-line voltages to the instruments and sensing relay.	FGT and MDID malfunction. MDID response sluggish.	Indication for line voltage values below 17.4 volts average.

Notes: Turn and bank indicator gyro is considered having response equivalent to gyro-horizon indicator.
    * Resistance ranges between 1.5 and 2.75 ohms for those fault conditions which result in an indication.
    GHID - Gyro horizon indicator.
    FGT - Flux-Gate transmitter.
    MDID - Master direction indicator.
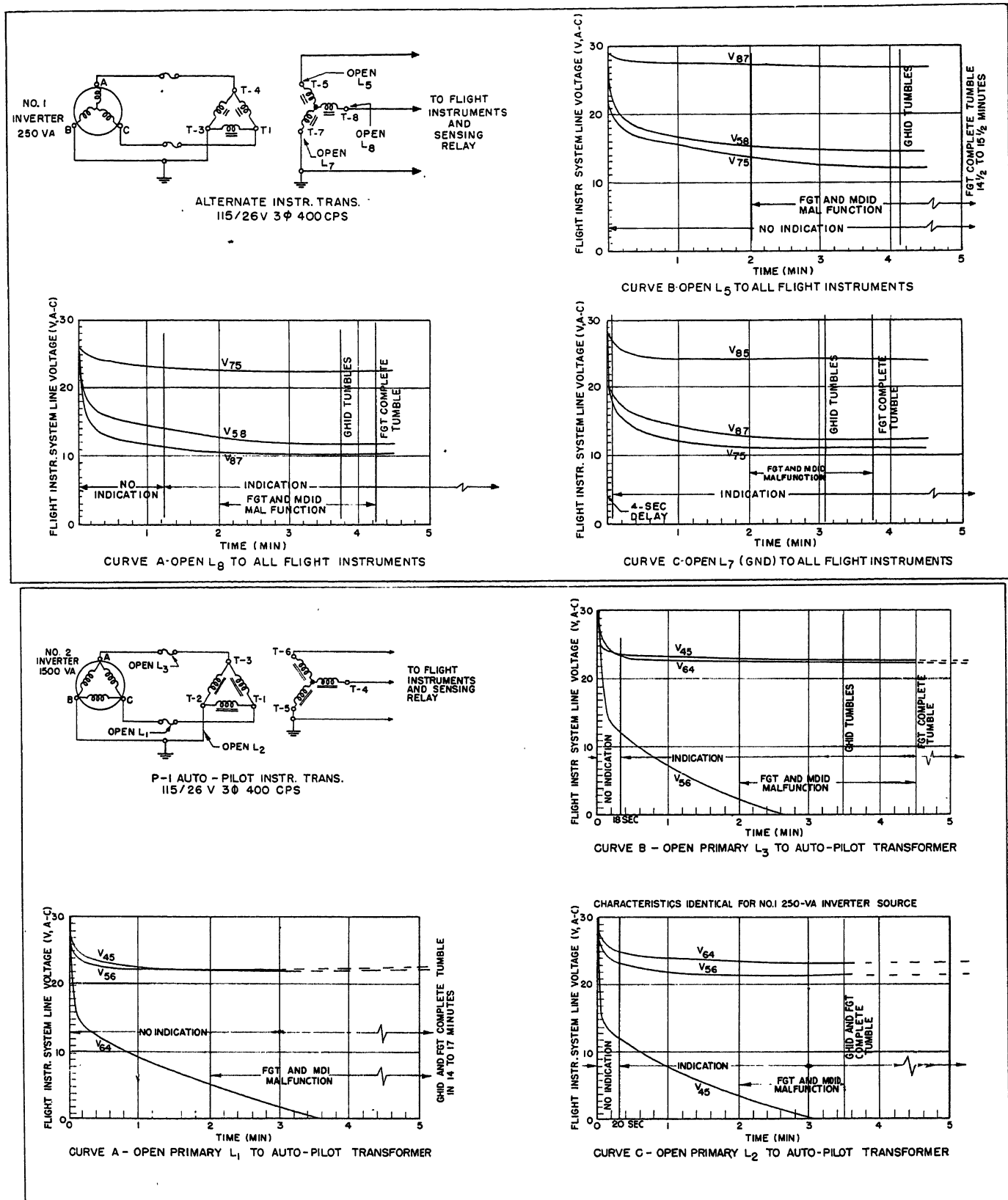
Fig. 8. Voltage decrease during open-circuit faults in the system shown in Fig. 4 and response of a 2-phase relay

carefully to avoid mechanical unreliability inherent in certain makes and models.

7. Corrective capacitance should not be used.

The present *MIL-E-7016* specification should be expanded to provide means of handling unbalanced transformations between a 115/200-volt wye source and a delta 115- or 26-volt distribution. This specification should also include a more accurate method of handling instrument power system analysis.

Better instrument power system designs would be furthered if, in a given "black box" load containing both 3-phase

balanced load and single-phase load, the power leads for each are terminated separately to allow the system designer the option of selecting that phase connection for the single-phase load which would best suit balancing the 3-phase system. Future developments of aircraft instruments should be guided so that they present a 3-phase balanced load directly to a 115/200-volt wye power source.

Development should be considered to provide fuses directly applicable to 3-phase circuits. These fuses should be capable of opening two lines or three lines when any individual line is overloaded sufficiently to open a fuse in that line.

There is a strong need for a fault detection device which can determine adequacy of a power source to deliver instrument power. This device should have elements of selectivity so that faults within the instruments do not create erroneous indications and also will not be affected by power feedback through the load side. When such a device is developed, further attention should then be devoted to completely automatic operation of the power transfer systems for aircraft instruments.

Summary

Design of electric systems for aircraft instruments has not been given the attention deserving of its critical safety-of-flight nature. Design aspects are brought to light which have been given only cursory attention in the past, resulting in decreased reliability. Information is provided to emphasize those design aspects and recommendations are offered toward improving future designs of aircraft instrument power systems.

References

1. EMERGENCY ELECTRICAL POWER FOR AIRCRAFT. Technical Report No. 6608, U.S. Air Force, Washington, D. C., Oct. 1952.

2. METHOD FOR AIRCRAFT AC ELECTRICAL LOAD ANALYSIS. Military Specification MIL-E-7016, Washington, D. C., Sept. 13, 1950.

3. CHARACTERISTICS OF AIRCRAFT ELECTRIC POWER. Military Specification MIL-E-7894, Amendment 1, Washington, D. C., Aug. 14, 1952.

4. CIRCUIT PROTECTION OF INSTRUMENT SYSTEMS POWERED BY 115 VOLT, 3 PHASE, 400 CYCLE INVERTERS. Technical Memorandum Report WCLE-53-251, Wright Air Development Center, Wright-Patterson Air Force Base, Ohio, Sept. 16, 1953.

5. PHASE REPORT No. 2, DEVELOPMENT OF IMPROVED ELECTRIC SYSTEM FOR FLIGHT INSTRUMENTS IN NAVY MODEL PBM-5A AIRCRAFT. Report No. NADC-EL-53131, Naval Air Development Center, Johnsville, Pa., Nov. 17, 1953.

6. ELECTRICAL PROTECTION OF AIRCRAFT FLIGHT INSTRUMENTS. Report . No. NADC-EL-167-50, Ibid., June 28, 1950.

7. VIBRATION TESTS OF INSTRUMENT SYSTEM FUSES. Technical Memorandum Report WCLE-53-296, Wright Air Development Center, Wright-Patterson Air Force Base, Ohio, Sept. 25, 1953.

8. EVALUATION OF THE HARTMAN ELECTRICAL MANUFACTURING COMPANY AVR-765D FAILURE SENSING RELAY FOR USE IN 26V AC FLIGHT INSTRUMENT SYSTEMS. Phase Report No. NADC-EL-5374, Naval Air Development Center, Johnsville, Pa., Sept. 25, 1953.

9. FINAL REPORT, DEVELOPMENT OF IMPROVED ELECTRIC SYSTEM FOR FLIGHT INSTRUMENTS IN NAVY MODEL P4M-1 AIRCRAFT. Report No. NADC-EL-5447, Ibid., July 24, 1954.

10. ELECTRIC POWER SYSTEM FOR A-C FLIGHT INSTRUMENTS. Dwg. No. R-157-B, Bureau of Aeronautics, Washington, D. C., Sept. 1953, sheets 1 and 2.

11. ENGINEERING INVESTIGATION OF GENERAL ELECTRIC COMPANY PHASE ADAPTERS. Report No. ADS-EL-129-49, Naval Air Development Center, Johnsville, Pa., July 20, 1949.

12. DESIGN STUDY FOR AIRCRAFT EMERGENCY AND INDEPENDENT ELECTRIC SYSTEM. Technical Memorandum Report No. ECLE-54-89, Wright Air Development Center, Wright-Patterson Air Force Base, Ohio, Dec. 31, 1954.

13. FINAL REPORT, DEVELOPMENT OF AN IMPROVED A-C FLIGHT INSTRUMENT AND COMPLETE D-C ELECTRIC SYSTEM FOR NAVY MODEL R4Q-1 AND 2 AIRCRAFT. Report No. NADC-EL-5534, Naval Air Development Center, Johnsville, Pa.

Directional Relays Provide Differential-Type Protection on Large Industrial Plant Power System

M. M. GILBERT
ASSOCIATE MEMBER AIEE

R. N. BELL
ASSOCIATE MEMBER AIEE

SEVERAL years ago it became necessary to revise the primary distribution system in a large chemical plant. This plant was approximately 25 years old and had undergone a rapid expansion during this entire period. The type of electric distribution system which had developed is typical of many other large chemical plants which have expanded over comparable periods of time. The original plant electric system consisted of 2,300-volt generation and primary distribution. Load increases associated with later plant expansions resulted in the introduction of 11,000-volt generation and primary distribution and the installation of an 11,000-volt interconnection with the local public utility. Bus tie reactors had been installed to keep the generator bus fault levels down as

additional power capacity was added to the system.

The primary objectives in the program of revising the distribution system were to reduce voltage variations resulting from utility swings and load changes, provide selective high-speed relaying, and further reduce the fault levels on the two 2,300-volt generator busses. Other important considerations were to make the maximum use of existing equipment and provide for installing equipment with a minimum amount of shutdown time to minimize losses in plant production.

Plant Primary Power System

The peak load of this plant at the time the study was begun was approximately 31,000 kva with the majority of the plant

electric load being supplied from five 6,250-kva turbine generators in the powerhouse. Two of these generators are 2,300-volt units and three are 11,000-volt units. In recent years, additional electrical capacity had been provided by the installation of an 11,000-volt interconnection with the local utility company. The five generators are connected to four main generator busses. Two of these busses operate at 2,300 volts and two at 11,000 volts. A duplex reactor is used to interconnect the two 2,300-volt busses and a second duplex reactor is used to interconnect the two 11,000-volt busses. An 11,000- to 2,300-volt bus tie transformer is used to interconnect the 2,300-volt and 11,000-volt duplex reactors. The 11,000-volt feeder from the public utility was connected to the primary side of this bus tie transformer. Fig. 1 shows a single-line diagram of this power system.

The 2,300-volt busses which originally supplied all of the plant loads now

primarily supply the auxiliaries in the powerhouse. The plant production facilities are now supplied by underground primary cable feeders from the two 11,000-volt busses.

The 2,300-volt switchgear consists of relatively old oil circuit breakers installed in concrete cell structures. The 11,000-volt bus structures consist of more recent drawout-type oil circuit breakers in metal-clad switchgear.

At the time the study was initiated, approximately 2,000 kva of additional electric load were being installed in this plant. Operating experience with the existing primary system arrangement had indicated that there were several undesirable conditions which would have to be corrected to provide the dependable electric power supply that was required for the increased plant loads. The duplex reactors which were installed to reduce fault levels on the generator busses and also introduced appreciable voltage differentials between the four generator busses for various operating conditions. These voltage differentials were most pronounced during the periods when one of the plant generators was out of service and all of the load on the bus was supplied through one leg of the duplex reactor. Maintenance schedules for boilers, turbines, and generators along with variations in process steam requirements resulted in the plant having one of the five generators out of service for as much as 6 months of every year.

The public utility substation for this plant was connected to the utility company's 110-kv transmission system at a point adjacent to one of their large generating stations. The voltage level at this generating station was changed in increments every day in order to keep their over-all system voltage levels within prescribed limits with normal daily variations in system load. The changes in the public utility generating station voltage produced similar voltage variations in the power supply to this plant. These public utility variations further added to voltage conditions inherent with relatively large reactance between powerhouse busses and resulted in bus voltage variations for normal operating conditions that were greater than could be tolerated. The maximum voltage variation on a generator bus for normal operating conditions was approximately 11 per cent.

In addition to the problem of voltage variations, all of the relaying associated with the primary distribution system needed revision. The two duplex reactors and the bus tie transformer were
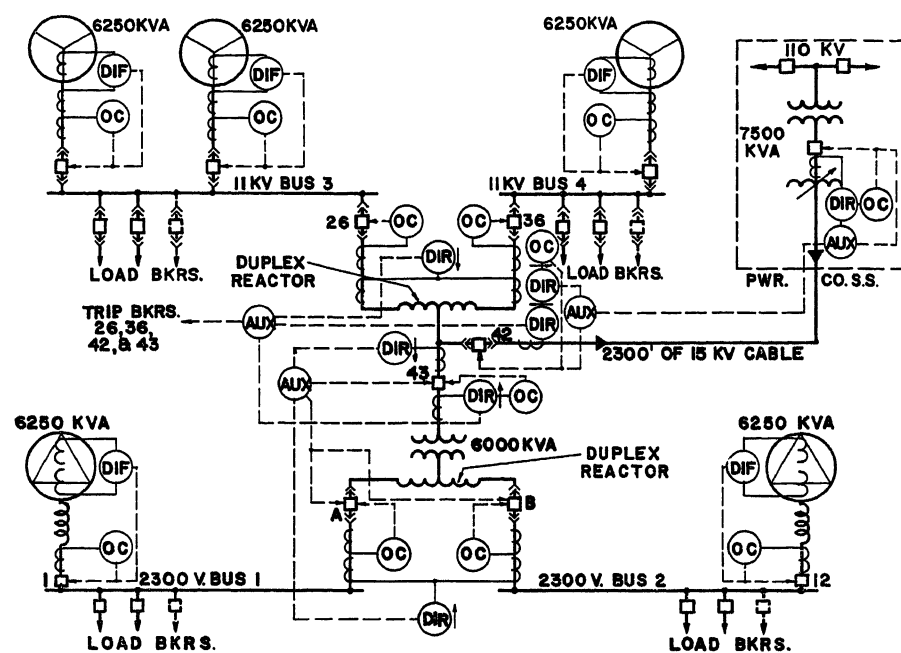
Fig. 1. Plant primary distribution system

protected by a differential relaying scheme in which current transformers associated with the five breakers surrounding this section of the system were all paralleled into a single set of differential relays. The performance of the current transformers associated with most of these breakers was very poor, and as a result the existing differential relaying scheme was subject to false operation for primary feeder faults. Several cases of misoperations of this differential relaying scheme had been reported. These false operations of the relay protection were costly because the load on one of the 11,000-volt busses exceeded the generation and therefore the entire load on this bus was lost whenever it was separated from the rest of the primary system. The existing relaying on the 2,300-volt generators consisted of differential relays around each of the 3 phases of the delta-connected generators. No protection existed for approximately 50 feet of 2,300-volt bus between the generators and the 2,300-volt switchgear. There was no existing relay protection for bus faults on the four generator busses. All of the over-current relays on the feeder breakers were definite minimum-time-type characteristics. In most cases, these definite minimum-time relays could not be set to provide the desired co-ordination with other existing overcurrent devices.

Primary Distribution Study

The first step in the study was to make some preliminary fault current calculations on the existing primary distribution system. These calculations indicated that the fault currents on the two 2,300-volt busses were approximately 160 per cent of the circuit-breaker ratings. On the basis of this information, a power study was begun with the objective of providing a primary power system which would eliminate the problem of interrupting duty on the 2,300-volt breakers and which would reduce voltage variations on the powerhouse busses to acceptable limits. Approximately 12 different system arrangements were developed. Each of these systems was analyzed for required investment, operating advantages, maximum use of existing equipment, and required plant shutdown time. As a result of this evaluation, four of the system arrangements were selected as having sufficient merit to warrant more careful consideration and comparison. An a-c calculating board study was made to obtain data for comparing each of the primary system arrangements. Each of the schemes was set up on the a-c calculating board and fault current and complete load and voltage studies were conducted. Primary system modifications which were considered ranged from minor revisions to the existing primary system to arrangements which involved converting all of the plant generation to 11,000 volts.

As a result of this a-c calculating board study along with economic evaluation of all of the schemes that were considered, it was determined that the voltage problems could be adequately solved at a minimum cost by the installa-
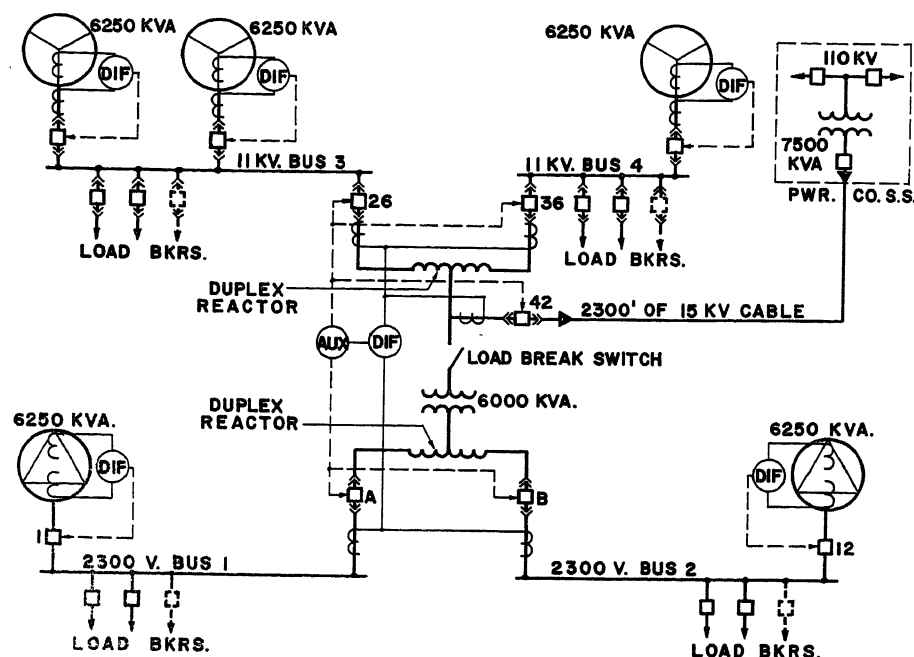
Fig. 2. Zones of protection for directional relaying

tion of a voltage regulator in the incoming feeder from the local power company. In addition, it was decided to install a new 11,000-volt breaker in place of the load break switch on the primary side of the bus tie transformer to provide more operating flexibility and to simplify the problem of relaying. The a-c calculating board study further showed that, with the installation of the voltage regulator to control reactive current flow, it would be possible to install current-limiting reactors in the 2,300-volt generator leads to reduce the fault level within the interrupting rating of existing breakers. Fig. 2 shows the revised primary distribution system incorporating these changes.

High-Speed Relay Protection for Powerhouse Bus Interconnections and 11,000-Volt Feeder to Power Company Substation

After having determined the basic modifications required in the plant primary distribution system, the next problem was to design modern high-speed relaying for the revised primary system arrangement. It was decided that the existing differential relaying around the generator bus interconnections should be removed both because of the possibility of false operation and also because a new 11,000-volt breaker was to be installed in the center of the existing differential relay zone. It was also concluded that high-speed relay protection was required for the powerhouse bus interconnections and for the 11,000-volt feeder to the

power company substation. The interconnections between the powerhouse busses actually involved three zones of relay protection. One zone included two sections of 2,300-volt metal-clad bus, the 2,300-volt duplex reactor, the outdoor bus tie transformer, and a relatively long 11,000-volt cable run. The second zone was a 4-terminal zone which contained the 11,000-volt duplex reactor and cable connections to the four 11,000-volt circuit breakers. The third zone of high-speed relay protection consisted of a 2,300-foot 11,000-volt cable feeder to the public utility substation and the voltage regulator. These three zones of relay protection are referred to as zones 1, 2, and 3 respectively, and are shown in Fig. 2.

The following five types of relaying were considered for this primary distribution system:

1. Restraint-type differential relays.
2. Differential relays in conjunction with air-core current transformer.
3. Pilot-wire relays.
4. Directional relays supervising auxiliary relays.
5. Voltage restraint directional relays.

Considerable operating experience had been accumulated with restraint-type differential relays and they would be normally given first consideration in applications similar to relay zones 1 and 2. However, detailed analysis of the existing current transformers along with careful study of various differential relay characteristics uncovered a number of obstacles to the use of this type of relay-

ing on this particular installation. The performance of the current transformers was extremely poor, and the possibility of false operation had to be carefully considered in all differential relay applications. Differences in current transformers ratios on the circuit breakers in a given zone of relay protection was another problem associated with differential relay protection.

One type of restraint differential relaying considered could not be applied to this plant primary system because of the current transformer accuracy and also because the existing fault levels and current transformer ratios produced relay fault currents in excess of the allowable maximum as recommended by the manufacturer. Other types of differential relaying were investigated. In all cases, the differential relay protection schemes for the 4-terminal relay zone 2 were relatively complicated. The complications encountered along with the possibility of false operations which might result from the poor performance of current transformers resulted in a decision against the use of this type of relaying for the powerhouse bus interconnections.

Another scheme which was considered was the use of air-core current transformers with associated differential relays. This type of relaying scheme involved the installation of air-core current transformers on each of the circuit breakers in the interconnections between the powerhouse busses. Relatively long outages would have been required to install these air-core current transformers. This scheme was discarded because the manufacturer could not supply current transformers of this type to go into the existing switchgear units in the powerhouse.

Pilot-wire relaying could have been applied on all three of the zones on which high-speed relay protection was desired. This type of relaying is the conventional method of providing high-speed protection for an important system tie circuit application similar to relay zone 3. The use of pilot-wire relays was ruled out because of their relatively high cost and also because it was felt that the plant electrical maintenance people could not satisfactorily maintain the relatively complex pilot-wire relays.

The restraint differential, the air-core current transformer, and the pilot-wire relay schemes are all fundamentally straight differential-type relaying. The relatively poor accuracy of the existing current transformers which we were attempting to utilize represented a serious

threat to the satisfactory application of any differential relaying scheme. For this reason, it was logical to attempt to develop a relaying scheme which did not involve the basic fundamentals of differential relaying which would provide the same high-speed, selective relay protection normally associated with differential relays.

Attempts to divert from straight differential relaying resulted in the consideration of two schemes which made use of directional relays to provide differential-type protection. Basically these directional relay schemes involve the installation of 3-phase directional relays on each of the circuit breakers associated with sources of power supplying a zone of relay protection. Each directional relay is connected so that the relay contacts close for power flow into the protected relay zone. All directional relay contacts are connected in series in a d-c control circuit. A fault in the protected zone results in the closing of all directional relay contacts energizing the d-c control circuit, thereby tripping all of the circuit breakers surrounding the relay zone.

One type of directional relaying which was evaluated involved the use of sensitive 3-phase directional relays. The sensitivity of these relays made it necessary to add a certain amount of time delay to the relay scheme. This time delay was introduced by the use of a d-c auxiliary relay in conjunction with each directional relay. These auxiliary relays introduced a 2 to 3-cycle time delay to prevent misoperation of this relaying for transient conditions. This relay scheme was workable but was more complicated and slower in operation than the directional scheme which was finally selected for this application.

The other directional scheme which was developed for this application was based on the use of 3-phase voltage restrained directional relays, designated type HV-3. The HV-3 relay was designed so that a relatively large amount of power is required to close the relay contacts with rated voltage on the relay. However, as the voltage is reduced, the relays become very sensitive. With this relay characteristic, all of the relay contacts in the directional relay scheme would be normally held open. With this condition, it was decided that it would not be necessary to install auxiliary relays to provide additional time delay to prevent false operation due to transient load swings resulting from system switching. Contacts of the HV-3 relays in a given protected zone were connected in

series with the d-c coil of a high-speed, manual-reset, latch-type auxiliary relay. The auxiliary relay had one set of contacts for each of the circuit breakers associated with the zone of relay protection. These contacts were wired in the d-c trip circuits of each of the breakers.

This latter scheme was selected for providing high-speed relay protection for power house bus interconnections and the 11,000-volt feeder to the power company substation. Reasons for the selection of this scheme were that engineering investigation indicated it would provide the type of protection that was required and it was the simplest and lowest cost of any relaying schemes which were considered for this application.

Details of Directional Relay Protection

The HV-3 directional type of relaying was applied to three sections of the primary power system in this plant. Fig. 3 shows details of each of these three zones of relay protection.

The 2,300-volt duplex reactor plus the 11,000-volt to 2,300-volt bus tie transformer was included in one zone of the directional relay protection. This section was designated as relay zone 1 and contained all of the primary system between 2,300-volt circuit breakers A and B and 11,000-volt breaker 43. To provide the directional relay protection for zone 1, one HV-3 relay was installed for breakers A and B and a similar relay for 11,000-volt breaker 43. By paralleling the current transformer on 2,300-volt breakers

A and B, it was possible to use a single HV-3 relay for both of these breakers. The HV-3 relay associated with 2,300-volt breakers A and B was connected so that its contacts closed for power flow toward the 11,000-volt system. The HV-3 relay on 11,000-volt breaker 43 was connected to close its contacts for power flow toward the 2,300-volt system.

The contacts on each of the two HV-3 relays in relay zone 1 were connected in series with a 125-volt d-c supply to the coil of a high-speed, manual-reset, latch-type auxiliary relay. A type-W.L. auxiliary relay was used for this application. The tripping of the W.L. relay closes three sets of contacts. One of these auxiliary relay contacts is connected in the d-c trip circuit of each of the three circuit breakers associated with this relay zone. Fig. 4 shows a schematic diagram of the d-c portion of the relay protection for all three zones of protection.

The HV-3 relay protection was also installed on the 11,000-volt portion of interconnections between the four powerhouse busses. This represents a 4-terminal section between 11,000-volt breakers 26, 36, 42, and 43 which included the 11,000-volt duplex reactor. This 11,000-volt portion of the powerhouse bus interconnections was designated as relay zone 2. Three HV-3 relays were installed to provide the directional protection for this zone. By paralleling current transformers, a single directional relay was used for breakers 26 and 36. Additional HV-3 relays were installed on 11,000-volt breakers 42 and 43. Each of the directional relays was connected so that its

Fig. 3. Revised primary distribution system
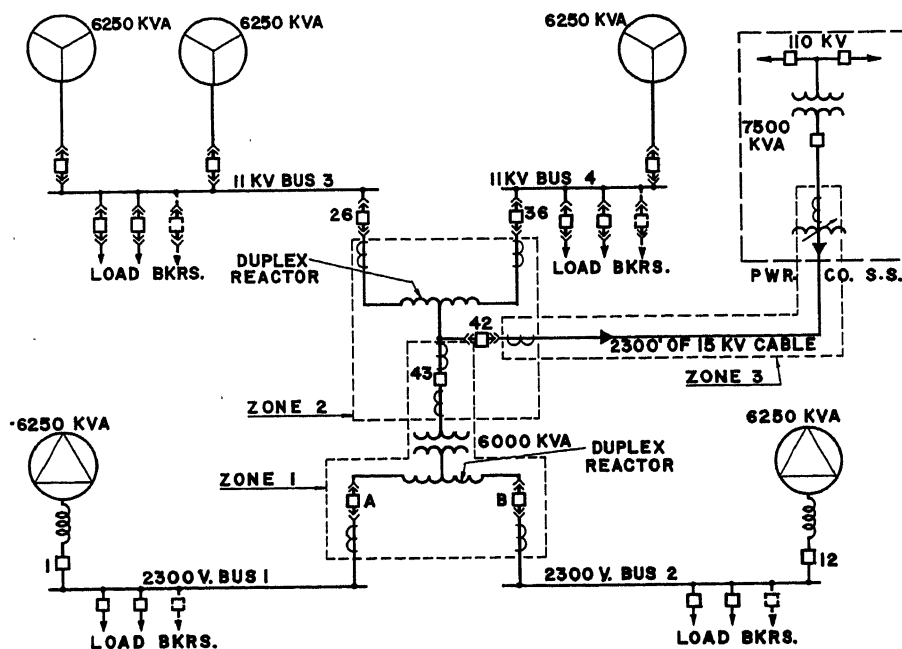
Gilbert, Bell—Directional Relay Protection
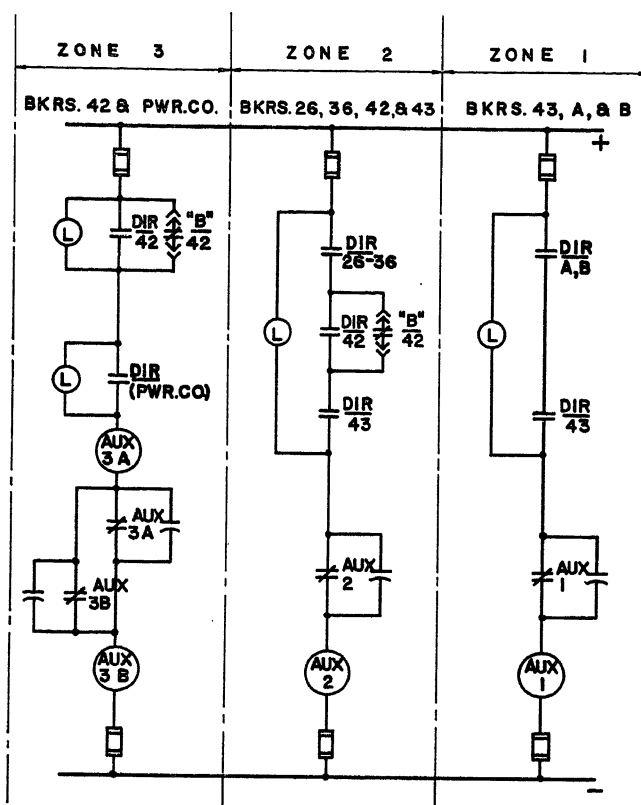
Fig. 4. D-c schematic of directional relay protection

contacts would close for power flow into the protected relay zone.

The contacts of the three HV-3 relays installed with this zone were connected in series with the 125-volt d-c supply to the coil of another W.L. auxiliary relay. Four sets of closing contacts were provided on this relay for tripping each of the four circuit breakers in zone 2. A b contact on circuit breaker 42 was connected in parallel with the contacts of the HV-3 relay on that breaker. This provided for keeping the high-speed directional relay protection in service when breaker 42 was open. This breaker was open during periods when all of the plant load could be supplied from powerhouse generators and no power was being purchased from the local public utility. Since this condition existed during certain periods of the year, it was decided that high-speed relay protection should be retained for this normal operating condition.

The 11,000-volt feeder to the power company represented the third section of the plant primary power system on which high-speed directional relaying was applied. This included approximately 2,300 feet of 15,000-volt cable and a step-type voltage regulator and is referred to as relay zone 3. Two HV-3 relays were installed for tl is zone of protection, one relay on 11,000-volt breaker 42 in the powerhouse and the other on the breaker in the power company substation. The HV-3 relay on breaker 42 is directional

for power flow toward the power company and the HV-3 relay on the plant company breaker is directional for power flow towards the plant powerhouse.

The d-c control circuit for this zone 3 relay consisted of two W.L. relays, one in the powerhouse and the other at the power company substation, and a cable pilot wire circuit between the two points. The contacts of the two HV-3 relays and the coils of the two W.L. switches were all connected in series across the 125-volt d-c supply at plant powerhouse. 33-volt coils were used for each of the W.L. relays. Signal lights were installed to supervise the pilot-wire circuit.

Special Considerations Associated With Directional Relay Protection

The relatively high amount of reactance between powerhouse busses in this plant made it necessary to consider voltage and current values for HV-3 directional relays for all types of faults within each of the three zones of relay protection. This was necessary to establish that there · was sufficient current to close the associated directional relays for all faults occurring in each of the three relay zones. Fault current studies made on the a-c calculating board provided the information necessary to check the operation of the relays for these conditions. A factor of safety to take care of tolerances in individual relay characteristics and to

allow for fault impedance limiting fault current below calculated values was included in the design of this directional relay protection. This factor of safety was obtained by designing the relay protection so that calculated values of fault current in the HV-3 relays were equal to three or more times the value required to just close the relay contacts with the corresponding voltages on the relay potential coils.

To meet the foregoing requirements for HV-3 relay on breaker 43 in zone 1 for faults adjacent to 2,300-volt breakers A and B, it was necessary to increase the sensitivity of this relay beyond that obtainable from the standard relay design. Increased sensitivity was obtained by reducing the amount of voltage restraint on the relay. The reduction in voltage restraint was accomplished by inserting resistors in series with restraint voltage coils on this relay. The standard relay requires 10 amperes at full voltage to close its contacts. By the installation of the resistors, the HV-3 relay on breaker 43 characteristic was changed so that only 5.0 amperes were required at full voltage to close the relay contacts. The 5.0 amperes of relay current correspond to 8,000 kva load through this 11,000-volt breaker 43. Since this is considerably above the maximum through this breaker, there was still sufficient restraint on the relay to keep its contacts open for all anticipated load conditions.

The possibility of magnetizing inrush currents to the voltage regulator in relay zone 3 causing false operation of the directional relay protection was also investigated. The b breaker contact on breaker 42 short-circuits the HV-3 relay contact associated with this breaker whenever this breaker is open. When the 11,000-volt cable to the power company substation is energized by closing the 11,000-volt breaker in the power company substation, the HV-3 relay on this breaker is subjected to magnetizing inrush to the voltage regulator and 11,000-volt cable circuit. Because of the b contact arrangement on breaker 42, it is necessary that the HV-3 relay on the power company substation breaker remain open during this period to prevent the directional relay scheme from immediately tripping the breaker. Calculations indicated that this magnetizing inrush was considerably below the current required to operate the HV-3 relays. Operating experience since the installation of this relaying scheme has proved that the magnetizing inrush has no effect on the directional relay scheme.

Considerable study was given to further

extending the practice of using auxiliary *b* breaker contacts to keep the high-speed directional relay protection in service when a breaker is open, as was done on breaker *42* in relay zones 1 and 3. The use of these auxiliary contacts on each of the circuit breakers in the directional relay zones appreciably complicated the relaying scheme and also introduced other problems which had to be solved by the installation of additional auxiliary relays. A thorough study of powerhouse operating conditions indicated that breaker *42* was the only breaker open during any appreciable amount of time during the year. For these reasons and also because it was firmly believed that the relay protection would be more reliable if it were kept as simple as possible, it was decided to eliminate the auxiliary *b* contact arrangement for all but breaker *42* in the three zones of relay protection.

Associated Relay Modernization Program

In addition to the directional relay protection, a number of other improvements were made in the relaying on the primary distribution system. The 2,300-volt generators were originally delta-connected and protected by two current transformers and a differential relay around each of the three generator windings. Approximately 50 feet of 2,300-volt bus running between each of these generators and the generator breaker was unprotected. By moving three of the existing generator current transformers into the circuit-breaker cell structure and making use of a special delta-to-wye differential connection, both the genera-

tor and the 2,300-volt bus were included in the differential protection.

Voltage-controlled overcurrent relays were installed on each of the five generators. Backup overcurrent relays were added to each of the powerhouse bus tie circuit breakers (breakers *A*, *B*, *26*, *36*, *42*, and *43*). These generator and bus tie breaker overcurrent relays provide protection for faults on each of the four generator busses. They also serve as backup protection for faults in the three zones of high-speed directional relay protection.

The overcurrent relays on most of the 2,300-volt feeder breakers had to be co-ordinated with large 2,300-volt motor fuses. Most of the overcurrent relays on the 11,000-volt feeder breakers had to be co-ordinated with the low-voltage air circuit breakers. The slope of the fuse and the low-voltage circuit breaker time-current characteristics is entirely different from the characteristic of the definite minimum-time overcurrent relay. For this reason, it was impossible in most cases to set the existing overcurrent relays to co-ordinate properly with the fuses or low-voltage circuit breakers.

For these reasons, all of the existing overcurrent relays are being replaced with either very inverse or extremely inverse types of relays. A number of starting tests were made indicating that the 2,300-volt motor fuses rating could be reduced below manufacturer's recommendations. The reduction of fuse ratings on certain of the larger motors also has helped in setting up the desired co-ordination between overcurrent devices.

Instantaneous trip units have been included in the new overcurrent relays on

the majority of 11,000-volt feeders and on certain of the relays being installed on the 2,300-volt feeder breakers. These instantaneous units provide for high-speed clearing of primary feeder faults.

Conclusions

The directional relay protection which is described in this paper has been in service at the plant for approximately 1 year. Operating experience is limited because to date there have been no faults within the three zones of directional relay protection. Satisfactory performance has been obtained to the extent that there have been no cases of false operation of the directional relay protection due to feeder faults or transients resulting from switching surges.

The fundamental principles used to develop the directional relay protection for this plant had been previously used by public utility relay engineers in similar relay schemes. Directional relay protection of this type could be effectively used under similar circumstances in other applications in electric utility and industrial power systems. However, it is not the intention to imply that this directional relaying can be universally substituted for differential relaying. The type of relay protection which was developed for this plant does demonstrate that conventional relaying schemes are not necessarily the best solution to the specialized relaying problems encountered in industrial plant power systems. Specialized relay schemes engineered to meet the specific requirements in the plant power distribution system should be given consideration for this type of application.

Discussion

Donn C. Achtenberg (The Detroit Edison Company, Detroit, Mich.): The use of directional relays to obtain differential protection appears to be a wise move when the available current transformers are such that normal differential connection gives unsatisfactory results.

At the Detroit Edison Company there is a scheme which concerns those positions whose definite time relays are to be replaced because their curves do not co-ordinate with fuses or low-voltage switchgear. With this system, the pickup of the definite minimum time relay is raised, giving sufficient clearance at all points of the curve. Then, to cover low-current faults, an instantaneous overcurrent relay is installed with the coil circuit in series with the coil of the definite time relay. The pickup of the instantaneous relay is set equal to that of the definite time relay before that pickup was raised. The contacts of the instantaneous relay are

connected to pick up an auxiliary relay, which is set to operate at about 1,500 cycles, and its contacts are connected in parallel with the definite time relay. There are many factors involved in determining whether this scheme may be used economically. If new inverse time relays fit the same panel drillings as the old definite time relays, then simple replacement may be the most economical solution. However, if the panels must be redrilled to fit the new relays anyway, it may prove more economical, if panel space permits, to add the less costly instantaneous relays plus the auxiliary relay. In the past, many definite time relays were replaced by inverse time relays just to allow for "cold-load" pickup after an extended outage, even if no co-ordination problem was involved. The great surplus of definite time relays which thereby accumulated prompted the development of this new scheme. It is now our practice on entirely new positions to use the new scheme, and a savings of about $50 per relay is realized.

Fig. 5 shows the trouble which arises from the attempt to obtain a definite time relay over a large fuse. Fig. 6 shows how the addition of the instantaneous relay plus an auxiliary alleviates this trouble. It may be noted from Fig. 6 that an inverse time relay set with a pickup equalling 600, and lever equalling 1, would follow approximately the same curve.

F. P. Brightman (General Electric Company, Schenectady, N. Y.): This is an interesting paper on modernizing relay protection. One of the problems often encountered is that of the old current transformers not having the necessary characteristics for use with the relays, a procedure which might first be considered based on general practice. When this happens, as it did in the case described, the interruption to production required to permit replacing the old current transformers usually results in discouragingly high costs. The problem in such cases is to find an acceptable substitute

Fig. 6. Trouble alleviated by addition of instantaneous relay plus auxiliary

Fig. 5. Trouble arising from attempting to obtain definite time relay over large fuse

for the so-called standard relaying systems that would otherwise be used.

For the case in question, the *HV-3* relay chosen, or its General Electric Company equivalent, type *CHP V*, would be a satisfactory solution to the problem since there will presumably always be a source of fault current to operate the directional relays in case of a short circuit inside the protected zone. That is important when directional relays are considered for use in this manner, because there must be sufficient short-circuit current either from generators or as pump-back from large motors to operate the relays at each of the terminals. Otherwise, the tripping circuit, consisting of all the relay contacts in series, will be incomplete and none of the breakers will be tripped.

The authors mention the substitution of very inverse time and relays and extremely inverse time relays for definite minimum-time relays ahead of fused starters for large motors. I suggest the importance of considerable caution, which the authors undoubtedly exercised, in applying extremely inverse time relays immediately ahead of large fuses. The possible difficulty is that, when an extremely inverse time relay is slowed down enough to ensure co-ordination with the fuse on high-magnitude fault current, it is very slow on moderate values of current because of the shape of its time-current curve.

The authors also mention that reducing the rating of motor-starter fuses below manufacturers' recommendations proved helpful in co-ordinating the fuses with relays ahead of them. This would occur because smaller fuses would have lower melting currents, and consequently the relays could have correspondingly lower settings. Again, caution is advised if such reduction are attempted; otherwise, unnecessary blowing of expensive fuses may result. Fuses for motor starters are selected on a conservative, but not overconservative, basis to ensure that they will more safely withstand many motor-starting and re-starting cycles. Unless a fuse is applied on such a basis, it is quite likely to be gradually weakened by the repeated heating and cooling cycles caused by motor-starting current and severe overloads, and will eventually fail unnecessarily. Even though the user was willing to accept some risk of such unnecessary blowing of fuses, it would be exceedingly difficult to determine the possibilities of such failures for a given reduction in fuse size below that recommended by the manufacturer. The fact that a given fuse will permit starting the motor a few times without blowing should not be interpreted as proof that that fuse is suitable for the proposed use over a long period.

George Steeb (Niagara Mohawk Power Corporation, Buffalo, N. Y.): Since 1941 a scheme for bus protection which can be applied in spite of differences in current transformer ratios has been installed in our Company on 16 busses of the 60-cycle system, and 23 busses of the 25-cycle system. The method, similar to that of the authors, has been used at six locations.

Until now, there have been eight different faults, that were cleared successfully, and no cases of incorrect operations.

No mention is made in the paper of how ground faults are detected and removed. Is the system free of transient fault voltages so that grounding becomes undesirable?

G. E. Grosser (Westinghouse Electric Corporation, Philadelphia, Pa.): The problem of revamping and adding to an existing system usually requires more study than is required on new plants. The authors are to be commended on their thorough study.

Duplex reactors, as used in this plant to keep interrupting duty on breakers within ratings, could no doubt be used more frequently to advantage. This type of reactor makes use of the close mutual coupling between the two sections so that a relatively low reactance is obtained in the circuit during normal operating conditions. Such a reactor does have higher reactance for a fault on one bus, which is highly desirable but which inserts high reactance in the circuit when a machine is out of service, thus presenting a voltage problem.

When an interconnection between an industrial generating plant and an electric utility is contemplated, the use of tap-changing under load on the tie transformer bank or a step regulator should be considered so that more control of the plant bus voltage may be obtained, a point which is brought out in the paper.

The authors' relaying might be called unconventional; at least it is unusual and would seem to meet the needs of this particular plant quite adequately and economically.

A Simple Method for Calculating the Time Response of a System to an Arbitrary Input

G. A. BIERNSON
ASSOCIATE MEMBER AIEE

THERE are a number of methods for calculating the response of a linear system to an arbitrary input which divide the input into a series of units so small that each can be approximated quite crudely, but they are all quite laborious and require a very high degree of accuracy of computation to achieve adequate accuracy in the calculated response. Even more important, they give very little insight into why the system responds the way it does, or how the system transfer function should be changed to improve the response.

On the other hand, there are two means of characterizing the time response of a system which are easily related to its frequency response, and which have proved very useful in design. These are the transient responses of the system to a step of position, a ramp, etc., and the error coefficients, or the related parameters, velocity constant, acceleration constant, etc. The transient responses describe the transient behavior of the system, in contrast to the steady-state behavior described by the error coefficients. In this paper it is shown how an input signal may be broken down into steady-state and transient components so that these two concepts can be combined to produce a simple means for calculating the response of a system to any input.

Although this technique can apply to any linear system, it is developed in terms of feedback control systems. The method of calculating the response to an arbitrary input is described, and a means of evaluating the accuracy of the calculated response is presented. The rules for approximating and constructing the response curve are then summarized in detail. Finally, to show how this method may be used quite simply in the design of feedback control systems, some general rules are presented for relating the shapes of the transient terms, which are used in calculating the response, to the major characteristics of the open-loop frequency response of the system. This paper is essentially a condensation of reference 1.

Calculation of System Response

SEPARATION OF INPUT INTO STEPS OF DERIVATIVES AND INTEGRALS

Most methods for calculating the response of a system to an arbitrary input require that the input be approximated by a series of impulses, or steps, straight lines, or parabolas. These units may be considered in a more general sense to represent steps of derivatives or steps of integrals of the input: a straight line is a step of the first derivative; a parabola is a step of the second derivative; and an impulse is a step of the first integral. By using a sufficient number of any one of these units, each method can be employed to approximate the input curve to any degree of accuracy.

In contrast, this paper shows how to break the input down into a sum of not just one type of unit but a number of types of units, using for each portion of the input curve that order of derivative step which best approximates that portion. For example, impulses should be used to approximate pulselike portions of the input, and parabolas should be used to approximate the portions which are nearly parabolic. To approximate the input in this manner generally requires much fewer approximating units than if only one type of unit is employed, and the operation of breaking the input down into the appropriate units is really quite easy. One merely has to differentiate (or integrate) the appropriate portions of the input curve, and at each derivative (or integral) approximate those portions which look like steps, as steps of that derivative (or integral), and subtract the approximated portions from the rest of the curve.

Thus, any input curve can be considered to represent a sum of steps of various derivatives and integrals. The steps of derivatives approximate the low-frequency portion of the input, and the steps of integrals approximate the high-frequency portions. To calculate the response to an arbitrary input, therefore, a general expression for the response of the system to a step of a derivative, and to a step of an integral, is necessary. Since the response to low-frequency inputs is usually of more interest in feedback-control applications, the response to a step of a derivative shall be considered first.

RESPONSE TO A STEP OF DERIVATIVE

The following analysis yields an expression for the response to a step of derivative, which contains a series of

steady-state coefficient terms and a transient term proportional to the size of the step.

To simplify the symbology, the nth time derivative of a function is designated as D^n. Thus

$$D^n x_i(t) \triangleq \frac{d^n x_i(t)}{dt^n} \tag{1}$$

Similarly, the mth time integral of a function is expressed as D^{-m}. Thus

$$D^{-m} x_i(t) \triangleq \underbrace{\int \int \int \int}_{m} dt^m x_i(t) \tag{2}$$

Derivation of Expansion

Assume that the input consists of a step of the nth derivative of magnitude A_n

$$D^n x_i(t) = A_n, \text{ for } t > 0 \tag{3}$$

Since A_n is a constant, the Laplace transform of this derivative is

$$\mathcal{L}[D^n x_i(t)] = \frac{A_n}{s} \tag{4}$$

The Laplace transform of the input is then

$$X_i(s) = \frac{1}{s^n} \mathcal{L}[D^n x_i(t)] = \frac{A_n}{s^{n+1}} \tag{5}$$

If the transfer function (X_o/X_i) is known, the transform of the response $X_o(s)$ may be obtained by

$$X_o(s) = \frac{X_o}{X_i} X_i(s) = \frac{X_o}{X_i} \frac{A_n}{s^{n+1}} \tag{6}$$

To take the inverse transform of $X_o(s)$, expand equation 6 in partial fractions as follows

$$X_o(s) = A_n \left(\frac{c_0}{s^{n+1}} + \frac{c_1}{s^n} + \ldots + \frac{c_n}{s} + T_n(s) \right) \tag{7}$$

The constants $c_0, c_1, \ldots$ and c_n are called the steady-state coefficients of the system, and $T_n(s)$ is called the nth-derivative transient term and represents the sum

$$T_n(s) = \sum_{l=1}^{L} \frac{K_{ln}}{s - s_l} \tag{8}$$

Each value s_l represents one of the L poles of the transfer function (X_o/X_i), and K_{ln} is the corresponding coefficient. To simplify the presentation of this analysis, it has been assumed that (X_o/X_i) has only single-order poles, but this restriction is not basic.

Now factor (A_n/s^{n+1}) from the steady-state coefficient portion of equation 7 and substitute $X_i(s)$ for (A_n/s^{n+1})

$$X_o(s) = X_i(s)(c_0 + c_1 s + c_2 s^2 + \ldots + c_n s^n) + A_n T_n(s) \tag{9}$$

Take the inverse transform, term by term, to obtain $x_o(t)$

$$x_o(t) = c_0 x_i(t) + c_1 D x_i(t) + c_2 D^2 x_i(t) + \ldots + c_n D^n x_i(t) + A_n T_n(t) \tag{10}$$

where $T_n(t)$ is the inverse transform of $T_n(s)$, and is equal to

$$T_n(t) = \sum_{l=1}^{L} K_{ln} \varepsilon^{s_l t} \tag{11}$$

Replace A_n by $D^n x_i(t)$ in equation 10

$$x_o(t) = c_0 x_i(t) + c_1 D x_i(t) + \ldots + [c_n + T_n(t)] D^n x_i(t) \tag{12}$$

If the input $x_i(t)$ is broken down into a sum of steps of derivatives, equation 12 shows that the response $x_o(t)$ to the input can be calculated as follows:

1. Multiply the smooth plot of $x_i(t)$ (i.e., with all step portions subtracted out) by c_0, the positional steady-state coefficient, to obtain the steady-state component of positional error $c_0 x_i(t)$.

2. Multiply the smooth plot of $D x_i(t)$ by c_1, the velocity steady-state coefficient, to obtain the steady-state velocity-component of error $c_1 D x_i(t)$.

3. Multiply the smooth plot of $D^2 x_i(t)$ by c_2, the acceleration steady-state coefficient, to obtain the steady-state acceleration component of error $c_2 D^2 x_i(t)$.

4. Continue this process for all the smooth derivative curves below the derivative where the input is finally approximated by a series of steps.

5. For each of the steps of the derivatives that are obtained, add a transient term given by the expression

$$(D^n x_i)[T_n(t) + c_n] \tag{13}$$

Finally, add graphically the plots obtained in steps 1 to 5 to get a complete plot of $x_o(t)$.

Calculation of Elements in Expansion

The methods for calculating the coefficients K_{ln} and c_n are given in reference 2. Since the poles s_l are of single order, K_{ln} is given by

$$K_{ln} = \left(\frac{(s - s_l)}{s^{n+1}} \frac{X_o}{X_i} \right)_{s = s_l} \tag{14}$$

as shown by equation 7, page 155, of reference 2. The poles at $s = 0$ are of multiple order, therefore c_n is given by

$$c_n = \left(\frac{1}{n!} \frac{d^n}{ds^n} \frac{X_o}{X_i} \right)_{s = 0} \tag{15}$$

as shown by equation 30, page 162, of reference 2.

Generally, the simplest way to compute the steady-state coefficients c_n

is by a long-division expansion of the transfer function $X_o(s)/X_i(s)$ of the system, as shown in Appendix I. The simplest way to obtain the set of transient coefficients K_{ln} is to calculate the coefficients for the input curve and from these values determine the coefficients for derivatives and integrals of the input. Setting n equal to zero in equation 14 gives

$$K_{l0} = \left(\frac{(s-s_l)}{s}\frac{X_o}{X_i}\right)_{s=s_l} \quad (16)$$

This gives the coefficients for a unit step-of-position input. Comparing equation 16 to equation 14 shows that the coefficient for the nth derivative K_{ln} is related to K_{l0} by

$$K_{ln} = \frac{K_{l0}}{(s_l)^n} \quad (17)$$

Equation 17 shows that as n is increased the coefficients for high-frequency poles (i.e., poles with large values of $|s_l|$) are decreased at a much faster rate than the coefficients for low-frequency poles (i.e., poles with small values of $|s_l|$). Consequently, if n is sufficiently large, the transient coefficient associated with the pole of the lowest frequency is much greater than all the other coefficients. Thus, for high-order derivatives, the transient terms are determined almost entirely by the lowest frequency pole or poles (X_o/X_i). If this pole is real, which often happens, the transient terms for the higher derivatives are simply exponentials.

In equation 12 the expression $[c_n + T_n(t)]$ is called the composite transient term for the nth derivative. To simplify the expression for this composite term, evaluate equation 12 at $t=0^+$. Since all of the derivatives of $x_i(t)$ below the

nth are zero at $t=0^+$, the equation becomes

$$x_o(0^+) = [c_n + T_n(0^+)]D^n x_i(0^+)$$
$$= A_n[c_n + T_n(0^+)] \quad (18)$$

If $n \geq 0$, i.e., if $x_i(t)$ is no more than a step discontinuity (not an impulse), $x_o(t)$ must be continuous, provided the system has a finite bandwidth. Since $x_o(t)$ is initially assumed zero, it thus must be zero at $t=0^+$. Setting equation 18 equal to zero gives

$$c_n = -T_n(0^+) = -\sum_{l=1}^{L} K_{ln}, \text{ for } n \geq 0 \quad (19)$$

Combine equations 11 and 19

$$[c_n + T_n(t)] = \sum_{l=1}^{L} K_{ln}(1-\epsilon^{s_l t}), \text{ for } n \geq 0 \quad (20)$$

Other expressions for the transient terms can be obtained as follows. Substituting equation 17 into equation 11 gives for the nth derivative transient

$$T_n(t) = \sum_{l=1}^{L} \frac{K_{l0}}{(s_l)^n} \epsilon^{s_l t} \quad (21)$$

By means of equation 21 it can be shown readily that

$$\frac{dT_n(t)}{dt} = T_{n-1}(t) \quad (22)$$

Although equation 21 does not apply if $T_n(s)$ has multiple-order poles, it can be proved that equation 22 holds for all transfer functions. Expressing equation 22 as a definite integral gives

$$[T_n(t) - T_n(0)] = \int_0^t T_{n-1}(t)dt \quad (23)$$

By equation 19, c_n is equal to $-T_n(0)$, for $n \geq 0$, so that

$$T_n(t) + c_n = \int_0^\infty T_{n-1}(t)dt, \text{ for } n \geq 0 \quad (24)$$

EXPANSION OF ERROR RESPONSE

For systems which behave as low-pass filters, in which the output x_o approaches the input x_i at low frequencies, the error x_e can be defined meaningfully as

$$x_e = x_i - x_o \quad (25)$$

This definition is also meaningful for feedback-control systems having unity feedback between the input x_i and the output x_o as shown in Fig. 1(A).

However, a more general definition for feedback-control applications is to consider the loop error x_e defined as

$$x_e = x_i - x_n \quad (26)$$

where x_n is the loop return signal for the loop input x_i, as shown in Fig. 1(B). Since the ideal operation of all feedback-control loops is to make the loop return signal x_n follow the loop input x_i (which it would do for infinite gain $|G(j\omega)|$), the loop error as defined by equation 26 is always meaningful because it defines the error between the ideal operation of the loop and the actual operation, although it does not necessarily define the error between the loop input and a particular loop output.

The error response is often considered for feedback-control applications because the loop return signal generally follows the loop input so closely that the most convenient way to evaluate the performance of the loop is to examine the loop error. Hence, an expansion for the error response shall be considered but, for simplicity, the definition of equation 25 shall be used. For feedback-control applications without unity feedback between x_o and x_i, the more general definition of equation 26 can be applied.
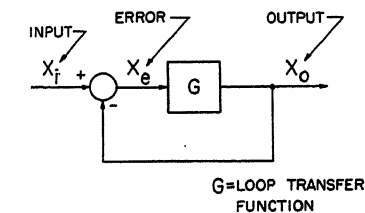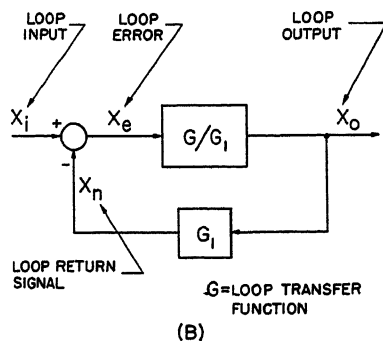
Fig. 1 (left). Definition of basic feedback loop variables
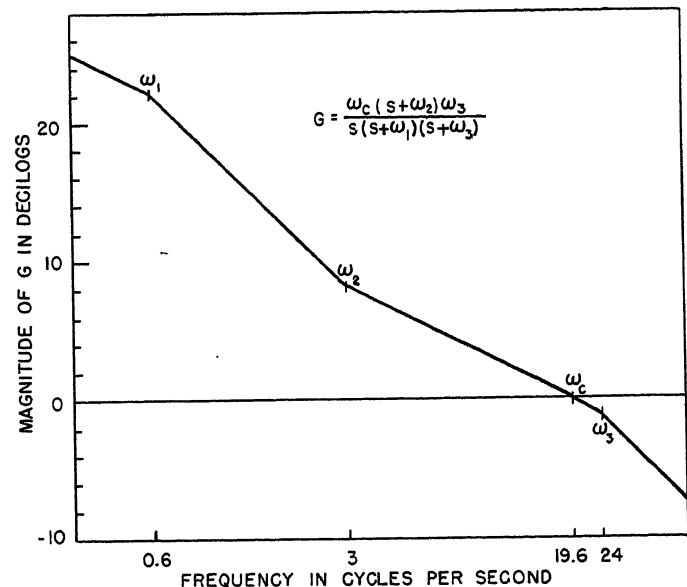
Fig. 2 (right). Magnitude asymptotes for illustrative system

Fig. 3. Position transient
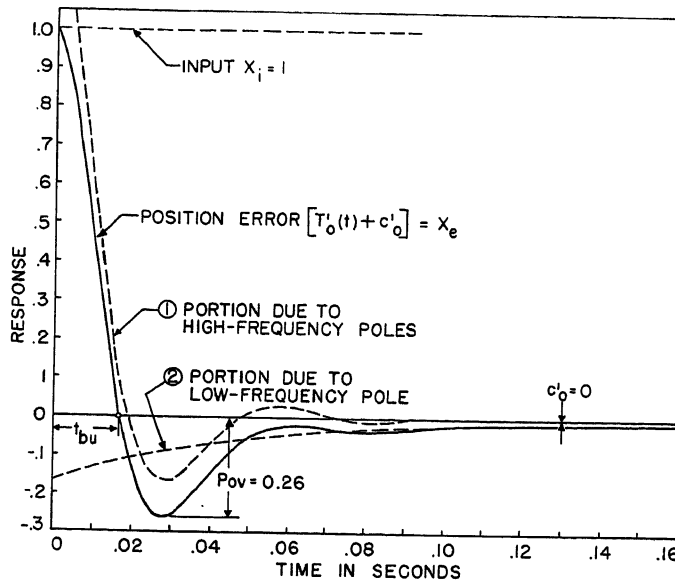
Fig. 4. Composite velocity transient

Substituting the expansion in equation 12 for $x_o(t)$ into equation 25 gives for the error response

$$x_e(t) = c_0' x_i(t) + c_1' D x_i(t) + c_2' D^2 x_i(t) + \ldots + [c_n' + T_n'(t)] D^n x_i(t) \quad (27)$$

where

$$c_0' = 1 - c_0 \quad (28)$$

$$c_n' = -c_n, \text{ for } n \geq 0 \quad (29)$$

$$T_n'(t) = -T_n(t) \quad (30)$$

The coefficients c_n' are termed the error coefficients[3] of the system. For a feedback-control system with a pure integration at zero frequency, the positional error coefficient c_0' is zero, and the velocity error coefficient c_1' is equal to the reciprocal of the velocity constant.[4] For a system with a double integration at zero frequency, c_1' is zero and the acceleration error coefficient c_2' is equal to the reciprocal of the acceleration constant.[4]

EXAMPLE

To illustrate this method, the error response of a typical feedback-control system to an arbitrary input shall be calculated.

The Illustrative System

The loop transfer function of the feedback-control system chosen for illustration is shown in Fig. 2. This loop transfer function is typical of a lead-network compensated 2-phase motor servomechanism operating at 60 cycles per second and employing a 5-watt motor such as a Diehl FPE 25–11. The system is adjusted for an M_p of 1.3. It has a velocity constant of 600 seconds[−1] and an asymptote crossover frequency

ω_c of 19.6 cycles per second or 123 radians per second. The vertical scale of Fig. 2 is in terms of the new general logarithmic unit, the decilog,[5] defined by

$$A \text{ in decilogs} = 10 \log_{10} A \quad (31)$$

The loop transfer function for the system is in terms of ω_c

$$G = \frac{X_o}{X_e} = \frac{\omega_c (s + \omega_2) \omega_3}{s(s + \omega_1)(s + \omega_2)}$$
$$= \frac{1.23 \omega_c^2 (s + 0.153 \omega_c)}{s(s + 0.0307 \omega_c)(s + 1.23 \omega_c)} \quad (32)$$

The closed-loop transfer functions are

$$\frac{X_o}{X_i} = \frac{G}{1+G} = \frac{1.23 \omega_c^2 (s + 0.153 \omega_c)}{(s + 0.175 \omega_c)(s^2 + 2\zeta \omega_n s + \omega_n^2)} \quad (33)$$

$$\frac{X_e}{X_i} = \frac{1}{1+G} = \frac{s(s + 0.0307 \omega_c)(s + 1.23 \omega_c)}{(s + 0.175 \omega_c)(s^2 + 2\zeta \omega_n s + \omega_n^2)} \quad (34)$$

where

$$\zeta = 0.522$$
$$\omega_n = 1.035 \omega_c$$

The system thus has a high-frequency pair of poles with the break frequency ω_n, and a low-frequency real pole with the break frequency $0.175 \omega_c$. These poles can be calculated readily by the techniques of reference 6.

The following error coefficients of the system can be calculated by a long-division expansion of X_o/X_i (or X_e/X_i), as shown in Appendix I.

position: $\quad c_0' = 0 \quad (35)$

velocity: $\quad c_1' = 1.628 \times 10^{-3} \text{ seconds} \quad (36)$

acceleration: $\quad c_2' = 3.53 \times 10^{-4} \text{ seconds}^2 \quad (37)$

rate of acceleration: $\quad c_3' = 1.715 \times 10^{-5} \text{ seconds}^3 \quad (38)$
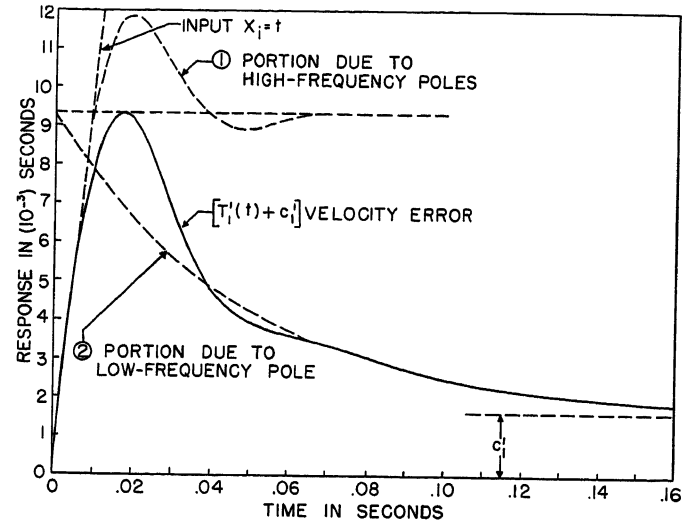
Composite Transient Terms

The first step in calculating the response of a system to an arbitrary input is to obtain its transient terms. The composite error-response transient terms $[c_n' + T_n'(t)]$ for the illustrative system are calculated in Appendix II and are examined in detail in the following.

Position Transient Term. The composite position transient term $[T_0'(t) + c_0']$ for the system is shown by the solid curve in Fig. 3, and represents the error response of the system for a unit-step-of-position input. Since the position error coefficient c_0' is zero, this is a plot of $T_0'(t)$. If the system did not have a pure integration in the loop, the plot would not go to zero at infinite time, and the final value of the plot would be c_0'. The broken-line curve x_i is a plot of the unit-step input to the system. The output response x_o can be obtained by subtracting the solid curve (error response x_e) from this step-input curve x_i.

The dashed curves of Fig. 3 show the separate effects of the two sets of closed-loop poles, curve 1 showing the component caused by the high-frequency complex pair of poles, and curve 2 showing the component caused by the low-frequency real pole. Note that the low-frequency real pole does not affect the build-up time t_{bu} appreciably, but it does have a pronounced effect upon the peak overshoot.

Velocity Transient Term. The solid curve x_e in Fig. 4 is a plot of the composite velocity transient term $[T_1'(t) + c_1']$, and represents the error response of the system for a unit-ramp input (step of velocity). Since $T_1'(t)$ becomes zero as t approaches infinity, the final value of this curve is the velocity error coefficient c_1'. The broken-line curve x_i

Biernson—Method for Calculating the Time Response of a System

represents the ramp input. For values of time before 0.01 second, there is very little motion of the output, hence the error x_e is approximately equal to the input x_i. At later values of time the output builds up, reducing the value of x_e, and eventually moves with the same velocity as the input but lagging it in position by the constant error c_1'. This plot of $[T_1'(t)+c_1']$ shows what the velocity-error coefficient c_1' means in terms of system performance: c_1' defines the velocity error for this system only for velocity components that have been maintained for a period greater than, say, 0.15 second; for velocity components maintained for a shorter period of time, the velocity error is much greater.

The dashed curve 1 shows the portion of the response caused by the high-frequency poles, and curve 2 shows the portion of the response caused by the low-frequency poles. The component caused by the high-frequency poles, curve 1, is displaced from the zero axis so that curve 1 approximates the total error response at short values of time; and the component caused by the low-frequency pole, curve 2, is displaced from the zero axis by the amount c_1', so that curve 2 approximates the error response at large values of time. Note that these two components, curves 1 and 2, are of roughly the same magnitude whereas, in the position transient term in Fig. 3, the component caused by the low-frequency pole is much smaller than that caused by the high-frequency pair of poles. This comparison bears out the statement made previously, following equation 17, that the effect of the low-frequency poles of a system upon the transient terms increases (relative to the high-frequency poles) as the order of differentiation is increased.

Acceleration Transient Term. The solid curve in Fig. 5 is a plot of the composite acceleration transient term for the error response $[T_2'(t)+c_2']$. The final value of this curve is the acceleration-error coefficient c_2'. This curve does not represent the complete error for a step-of-acceleration input, but merely that portion of error associated with acceleration (the acceleration error). For the acceleration to be a step, the velocity must increase as a ramp and the position must increase as a parabola. Since the position error coefficient c_0' is zero, there is no error associated with the parabolic position function, but there is an error associated with the ramp velocity function equal to the product of the velocity-error coefficient c_1' times the velocity. For a unit step of acceleration, this velocity error is

$$c_1't = 1.63 \times 10^{-3}t \tag{39}$$

A plot of this term is shown by the broken-line plot in Fig. 5 labeled velocity error. This velocity error is added to the acceleration error $[T_2'(t)+c_2']$ to yield the total error for a step of acceleration shown by a broken-line curve.

The input X_i, shown by a broken-line curve in Fig. 5, is the double integral of a unit step, and, hence, is a parabola given by the equation $(t^2/2)$. At values of time before $t=0.01$ second, the total error curve is essentially equal to the input X_i, which indicates that there is practically zero output motion up until that time. At greater values of time, the output moves to follow the input, but does so with an ever-increasing error. At very large values of time, most of this error is caused by the velocity component $c_1't$ but, for the values of time shown in this plot, the acceleration-error component is greater than the velocity component. Consequently, for many input signals to this system, acceleration errors can be quite significant with respect to velocity errors.

The/dashed curve 2 shows the com-

ponent of the acceleration transient term caused by the low-frequency pole. The small difference between curve 2 and the solid acceleration-transient curve represents the component caused by the high-frequency poles. The acceleration transient term is thus determined primarily by the low-frequency pole, the high-frequency pair of poles having little effect.

Rate-of-Acceleration Transient Term. The solid curve in Fig. 6 is a plot of the composite rate-of-acceleration transient term for the error response $[T_3'(t)+c_3']$. This term is negative, but for uniformity it is plotted in the positive direction in the figure, and the scale is labeled as negative. This term represents one of the three components of error which result from a unit-step of rate-of-acceleration input. The other two components are the acceleration error $c_2't$ which is equal to the acceleration error coefficient times the resultant ramp of acceleration, and the velocity error $c_1'(t^2/2)$ which is equal to the velocity-error coefficient times the resultant parabola of velocity. A plot of the acceleration error $c_2't$ is shown as a straight broken line. This error is positive, as indicated, and hence subtracts from the rate-of-acceleration error. It is important to note that the acceleration error is at all times greater in magnitude than the rate-of-acceleration error. Consequently, the rate-of-acceleration error is always of minor importance in determining the response of this system to an arbitrary input, and, hence, crude step approximations can be made of the rate-of-acceleration input curve.

The dashed curve 2 represents the component of the transient term caused by the low-frequency real pole, and the very small difference between this curve and the solid rate-of-acceleration transient curve represents the component caused by the pair of high-frequency poles. Since

Fig. 5. Composite acceleration transient

Fig. 6. Composite rate-of-acceleration transient

Fig. 7. Components of response to step of rate-of-acceleration

though curve 3, the velocity error, does) because curve 1 is in the opposite direction to the total error, and curve 2 is much greater than the total error for a significant period of time. However, the sum of these two, curve 4, does appear to be a legitimate component of error, because it is always less than the total error and is never negative. It is logical then to consider curve 4 as the corrected acceleration error. The acceleration error of curve 2 in a sense represents an overestimate of the "real" acceleration error, from which the rate-of-acceleration "correction" of curve 1 must be subtracted to yield the real corrected acceleration error of curve 4. The importance of this concept is that it shows the rate-of-acceleration transient term to be not an actual component of error, but merely a correction to the acceleration error, and, hence, of minor importance in determining the system response.

Breakdown of Arbitrary Input Into Steps of Its Derivatives

The arbitrary input position signal chosen for illustration is shown as curve 1 in Fig. 8. The system is initially at rest at an angle of zero degrees. At $t=0$ the position jumps discontinuously to -1.5 degrees, and then follows curve 1 in a smooth fashion as shown.

This input is broken down as follows. First, the step at $t=0$ is subtracted from the remainder of the curve and considered separately. Then the smooth portion of the position curve is differentiated, yielding the velocity curve 2.

At time $t=0.32$ second, the velocity

curve has a very quick, steep drop. Since most of this drop occurs in about 0.008 second, which is much less than the build-up time of the velocity transient term, the drop has approximately the effect of a step of velocity and, consequently, it is approximated by a step, as shown. If the velocity were differentiated in this region, it would produce a large, short pulse of acceleration having a magnitude of roughly 20,000 degrees per second-squared.

The smooth portion of the approximated velocity curve is differentiated to yield curve 3, the acceleration. At $t=0$ and $t=0.045$ second, the curve rises significantly faster than the acceleration transient and, hence, the acceleration curve is approximated by steps at these two points, as shown by the short dashed lines. The remaining portion of the acceleration curve is approximated by straight-line segments, as shown by the long broken lines. When the broken-line approximation is differentiated, it yields as the rate-of-acceleration, a series of steps, which are labeled curve 4. Thus, except for the portions neglected in the various approximations, the input is broken down into a series of steps of various derivatives of the input.

CALCULATION OF RESPONSE OF SYSTEM TO THE ARBITRARY INPUT

The plots of the composite transients shown in Figs. 3 to 6 are used to determine the response of the system to the arbitrary input function of Fig. 8. The approximated curves of velocity, acceleration, and rate-of-acceleration are multi-

the high-frequency poles have very little effect, this rate-of-acceleration transient term can be very accurately approximated by a simple exponential.

In Fig. 7 the three components of error for a step of rate-of-acceleration are shown: the rate-of-acceleration error, curve 1; the acceleration error, curve 2; and the velocity error, curve 3. The sum of curves 1 and 2 is curve 4, the acceleration-plus-rate-of-acceleration error. Adding curve 4 to the velocity error, curve 3, yields the total error, curve 5.

Curves 1 and 2 do not appear to be legitimate components of error (al-

Fig. 8. Illustrative arbitrary input and its derivatives

plied by the corresponding error coefficients to yield curves 1', 2', and 3' in Fig. 9. The velocity curve includes a single step approximation at $t=0.32$ second; the acceleration curve includes step approximations at $t=0.03$ second and $t=0.45$ second; and the rate-of-acceleration curve is entirely composed of step approximations. At each of the discontinuities of these curves, composite transient terms are added to produce the resultant curves of velocity error, acceleration error, and rate-of-acceleration error labeled 1, 2, and 3. The individual transient terms are shown as dashed curves; the complete error components as solid curves; and curves 1', 2', and 3', as dot-dashed curves in the regions where they do not correspond to the actual error curves 1, 2, and 3.

The largest transient in Fig. 9 occurs at 0.32 second and is caused by the approximate step of velocity at that point. The transient is obtained by multiplying the curve of $[T_1'(t)+c_1']$ in Fig. 4 by the step of velocity, which is -108 degrees per second. When this transient is added to curve 1' (which is $c_1'Dx_i(t)$), the resultant curve 1 cannot have a discontinuity because the discontinuity in curve 1' is c_1' times the step in velocity, which is also the final value of the transient.

The acceleration error curve has two transients, at 0.03 second, and at 0.45 second. Adding the two transients to the discontinuous curve 2' gives the continuous acceleration error curve 2. The rate-of-acceleration curve is entirely composed of steps, there being eight in all. For each of these steps there is a transient which adds to curve 3' to give the continuous rate-of-acceleration curve 3.

The three components of error in Fig. 9, curves 1, 2, and 3, are added together to form curve 1 in Fig. 10, which represents the total error caused by the derivatives of the input. Note that the vertical scale of Fig. 10 is half that of Fig. 9. Curve 2 in Fig. 9 is the error component caused by the -1.5-degree step of position which occurs at $t=0$. Curve 2 is obtained by multiplying the position transient term in Fig. 3 by the value of the step, -1.5 degrees. The total error response, curve 3, is obtained by adding curve 2 to curve 1.

Although discontinuities did not occur in the error-response components for the derivatives, there is a discontinuity in the position component, curve 2, and in the total error response, curve 3, at $t=0$. There must be a discontinuity in

the error response whenever there is a discontinuity in the input curve, because the output response must not have a discontinuity. If the output response were plotted instead of the error response, no discontinuity would occur.

RESPONSE TO A STEP OF INTEGRAL

If the input signal contains high-frequency components which rise and fall, like pulse, in periods significantly less than the rise time of the step response, it is very difficult to approximate them as steps. Instead, they should be integrated and approximated by impulses, doublets, etc. Thus, for these components the expansion for a step of an integral should be examined.

Assume that the mth integral of $x_i(t)$ is a step of magnitude A_{-m}, i.e.

$$D^{-m}x_i(t) = A_{-m}, \; t>0 \qquad (40)$$

The Laplace transform of the mth integral is

$$\mathcal{L}[D^{-m}x_i(t)] = A_{-m}/s \qquad (41)$$

Since this transform of the mth integral of x_i is equal to $X_i(s)/s^m$, the input $X_i(s)$ is

$$X_i(s) = A_{-m}s^{m-1} \qquad (42)$$

The output response for this input is

$$X_o(s) = \frac{X_o}{X_i}X_i = A_{-m}\frac{s^{m-1}X_o}{X_i} \qquad (43)$$

Equation 43 may be expanded as follows

$$X_o(s) = A_{-m}[a_0s^{m-1}+a_1s^{m-2}+\ldots+ \\ a_{m-1}+T_{-m}(s)] \qquad (44)$$

The transient term $T_{-m}(s)$ is obtained by replacing n by $-m$ in $T_n(s)$. The coefficients, a_0, a_1, ...a_m are called the initial value coefficients, and occur for the values of m at which $(s^{m-1}X_o/X_i)$ is an improper fraction. As shown in reference 2, page 154, the terms of equation 44 which contain these coeffi-

Fig. 9. Components of error response to arbitrary input
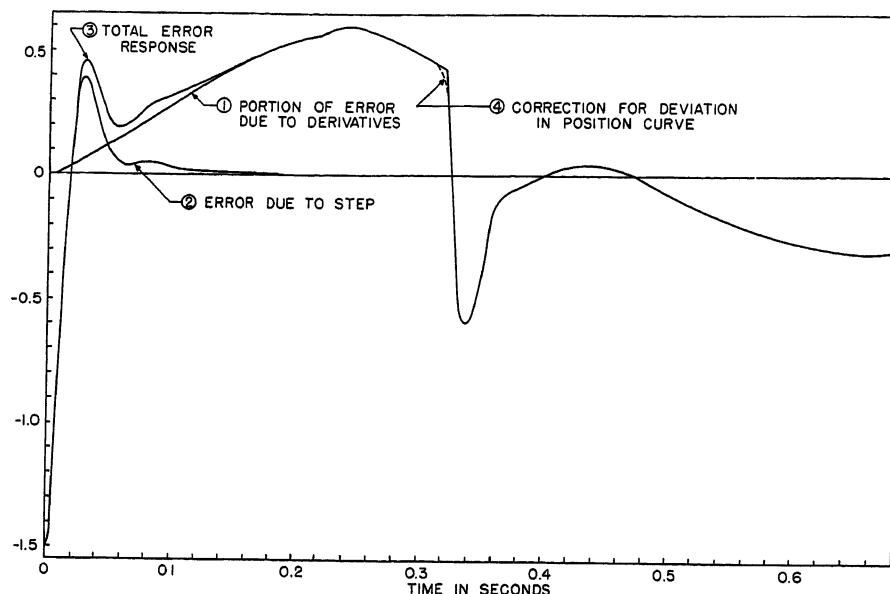
Fig. 10. Total error response to arbitrary input

cients produce impulses, doublets, triplets, etc., in the $x_o(t)$ response.

Substitute equation 42 into equation 44

$$\mathbf{X}_o(s) = a_0\mathbf{X}_i(s) + a_1\frac{\mathbf{X}_i(s)}{s} + \ldots +$$
$$a_{m-1}\frac{\mathbf{X}_i(s)}{s^{m-1}} + (A_{-m})\mathbf{T}_{-m}(s) \quad (45)$$

Take the inverse transform, and replace A_{-m} by $D^{-m}x_i(t)$

$$x_o(t) = a_0x_i(t) + a_1D^{-1}x_i(t) + \ldots +$$
$$a_{m-1}D^{-(m-1)}x_i(t) + [D^{-m}x_i(t)]\,T_{-m}(t) \quad (46)$$

Equation 46 shows that the output response to high-frequency portions of the input curve may be calculated as follows. Integrate each high-frequency portion successively and, at an appropriate integral, approximate the curve by steps. Multiply each integral curve, except the last, by the corresponding initial-value coefficient a_m and add the resultant curves to the calculated response. For each step add a transient

to the calculated response proportional to the magnitude of the step.

The initial-value coefficients can be calculated by replacing s by $(1/\lambda)$ in the transfer function $(\mathbf{X}_o/\mathbf{X}_i)$, and by employing the following equation which is similar in form to that used for calculating the steady-state coefficient c_n

$$a_m = \frac{1}{m!}\frac{d^m}{d\lambda^m}\left(\frac{\mathbf{X}_o}{\mathbf{X}_i}(1/\lambda)\right)_{\lambda=0} \quad (47)$$

The first coefficient a_0 is practically always zero, and very often a_1 is also zero. In any case, the lowest-order nonzero coefficient is equal to

$$a_m = \frac{s^m\mathbf{X}_o}{\mathbf{X}_i}\Big|_{s=\infty} = s^m G\Big|_{s=\infty} \quad (48)$$

The simplest way to calculate the initial-value coefficients, however, is to expand

the transfer function $(\mathbf{X}_o/\mathbf{X}_i)$ in long division, but with the order of the coefficients reversed from that used in calculating the steady-state coefficients, as is illustrated in Appendix I. The coefficients a_m are called initial-value coefficients because they represent the initial values of the corresponding integral transient terms, i.e.

$$a_m = T_{-m}(t)\big|_{t=0}^+ \quad (49)$$

For the illustrative system the initial-value coefficients are

position, impulse: $a_0 = a_1 = 0$ (50)

doublet: $a_2 = 1.85 \times 10^4$ seconds^{-2} (51)

triplet: $a_3 = 2.51 \times 10^6$ seconds^{-3} (52)

The integral transient terms of the illustrative system for a unit impulse (step of the first integral) and for a unit doublet (step of the second integral) are shown in Figs. 11 and 12. They are plots of $T_{-1}(t)$ and $T_{-2}(t)$ and represent the complete output response to a unit impulse and to a unit doublet respectively. The response of the system to a unit triplet would contain not only the triplet transient term $T_{-3}(t)$ but also an impulse of magnitude a_2, as shown by the expansion of equation 46. The impulse initial-value coefficient a_1 is the initial value of the impulse transient term $T_{-1}(t)$ which is zero, while the doublet initial-value coefficient a_2 is the initial value of the doublet transient term $T_{-2}(t)$.

Determination of Accuracy of Calculated Response

The calculation of the response to an arbitrary input requires that approxima-
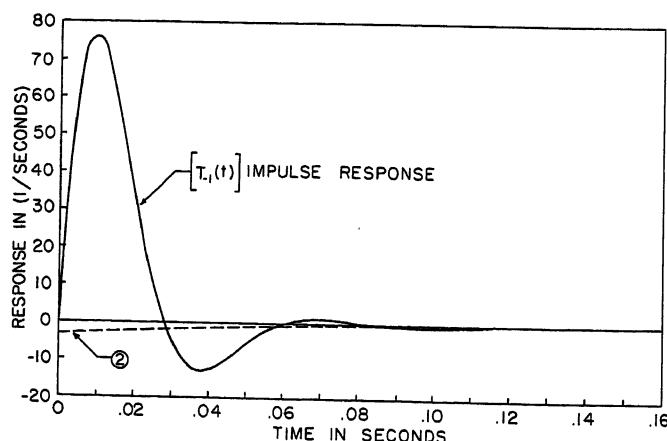
Fig. 11. Impulse transient
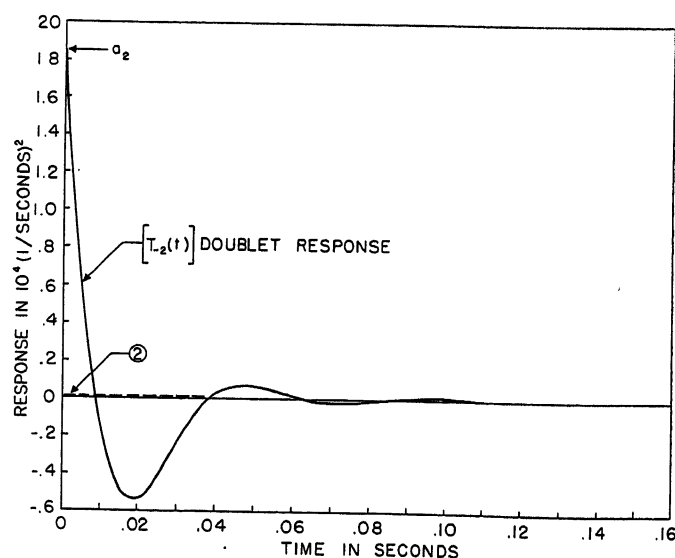
Fig. 12. Doublet transient

tions be made upon the input. The effect of these approximations upon the accuracy of the calculated response will now be considered.

Although it may be extremely difficult to calculate the exact effect upon the output of a deviation neglected in one of the approximations, it is quite easy to calculate an upper bound, which is at all time equal to or greater than the magnitude of the effect of that deviation on the output. This bound thus is a measure of the inaccuracy of the calculated response. If the approximation procedure is performed properly, this upper bound can be kept quite small with respect to the total error response, and, hence, the per-unit accuracy of the calculated response is guaranteed to be quite high.

An expression for the upper bound to the inaccuracy of an error-coefficient expansion was first derived by Bower.[7] However, the upper bound calculated herein is based upon an expression derived by Arthurs and Martin[8] because their expression yields a lower value for this upper bound.

The upper bound to the inaccuracy, corresponding to an approximation of a given derivative, is obtained by multiplying by the remainder coefficient of the derivative the maximum value of the magnitude of the deviation of the approximation from the actual curve. The remainder coefficient is calculated by a graphical construction upon the corresponding system transient term. A basic limitation of this inaccuracy bound is that it holds for all values of time, whereas the actual inaccuracy reduces to zero after the approximation has ceased. This limitation can be removed by generalizing the remainder coefficient to form the remainder transient which yields an inaccuracy bound that decays to zero following the end of an approximation. This remainder transient is also calculated graphically from the corresponding system transient term.

Derivation of Remainder Coefficient

The response to a deviation neglected in the nth derivative can be calculated as follows. $X_o(s)$ can be expressed as

$$X_o(s) = \frac{X_o}{X_i} X_i = \frac{X_o}{s^n X_i} s^n X_i \quad (53)$$

If the expression $s^n X_i$ represents the transform of the nth derivative deviation, equation 53 gives the transform of the corresponding portion of the response caused by that deviation. To evaluate equation 53, expand $(X_o/s^n X_i)$ as follows

$$\frac{X_o}{s^m X_i} = \frac{c_0}{s^n} + \frac{c_1}{s^{n-1}} + \ldots + \frac{c_{n-1}}{s} + T_{n-1}(s) \quad (54)$$

This expansion is complete provided that $(X_o/s^n X_i)$ is zero at $s = \infty$, which holds for $n \geq 0$. The expansion can be verified by considering the definitions for c_n and $T_n(s)$ given in equations 15 and 21. Combine equations 53 and 54

$$X_o(s) = c_0 X_i + c_1 s X_i + \ldots + c_{n-1} s^{n-1} X_i + s^n X_i T_{n-1}(s) \quad (55)$$

Take the inverse transform of equation 55

$$x_o(t) = c_0 X_i + c_1 D x_i(t) + \ldots + c_{n-1} D^{n-1} x_i(t) + D^n x_i(t) * T_{n-1}(t) \quad (56)$$

The last term of equation 56 is the real convolution of $D^n x_i(t)$ and $T_{n-1}(t)$, and is called the nth derivative remainder $\mathcal{R}_n(t)$. As shown in reference 2, page 228, this real convolution can be expressed as

$$\mathcal{R}_n(t) = \int_0^\infty d\tau T_{n-1}(\tau) D^n x_i(t - \tau) \quad (57)$$

This expression of equation 57 was first obtained by Arthurs and Martin[8] and represents the exact value for the remainder in an error-coefficient expansion. Thus, $x_o(t)$ can be expanded in the form

$$x_o(t) = c_0 x_i(t) + c_1 D x_i(t) + c_2 D^2 x_i(t) + \ldots + c_{n-1} D^{n-1} x_i(t) + \mathcal{R}_n(t) \quad (58)$$

Equation 58 shows that the effect on the calculated response of neglecting a deviation of the nth derivative is given by the remainder $\mathcal{R}_n(t)$ for that deviation. The steady-state coefficient terms of the expansion of equation 58 are automatically included in the calculated response when the procedure described previously is followed, so that it is only the term $\mathcal{R}_n(t)$ which is disregarded when the nth-derivative deviation is neglected.

Thus, the effect upon the error response of a deviation of the nth derivative could be calculated by substituting the equation for the deviation into the integral of equation 57, and by evaluating the integral. On the other hand, an exact calculation of $\mathcal{R}_n(t)$ for a given deviation curve would be extremely difficult, and it is not necessary because $\mathcal{R}_n(t)$ is quite small with respect to the total $x_e(t)$ function, when the approximations are performed properly. It is only necessary to determine a bound upon the magnitude of $\mathcal{R}_n(t)$ to give a measure of the accuracy of the calculated response.

To calculate this upper bound to $|\mathcal{R}_n(t)|$, note that

$$\left| \int_0^\infty d\tau T_{n-1}(\tau) D^n x_i(t - \tau) \right| \leq \int_0^\infty d\tau |T_{n-1}(\tau)| \cdot |D^n x_i(t - \tau)| \quad (59)$$

This equation merely states that the magnitude of an integral of a function can be no greater than the integral of the magnitude of the function. Equation 57 shows that the left-hand expression of this inequality is equal to $|\mathcal{R}_n(t)|$; hence

$$|\mathcal{R}_n(t)| \leq \int_0^\infty d\tau |T_{n-1}(\tau)| \cdot |D^n x_i(t - \tau)| \quad (60)$$

If the deviation from the nth derivative for which the remainder applies stays between the limits of $\pm N$ for all values of time, i.e., if

$$|D^n x_i(t - \tau)| \leq N \quad (61)$$

for all t and τ, an upper bound to $\mathcal{R}_n(t)$ is then

$$|\mathcal{R}_n(t)| \leq \int_0^\infty d\tau (N) |T_{n-1}(\tau)| = N \int_0^\infty d\tau |T_{n-1}(\tau)| \quad (62)$$

Define the remainder coefficient r_n as

$$r_n = \int_0^\infty d\tau |T_{n-1}(\tau)| \quad (63)$$

The upper bound to the inaccuracy caused by the approximation then may be expressed as

$$|\mathcal{R}_n(t)| \leq r_n N \quad (64)$$

where N is the maximum value of the magnitude of the deviation from the nth derivative over all values of time.

A corresponding expression can be obtained for a bound to the inaccuracy resulting from neglecting a deviation of the mth integral. If (X_i/s^m) represents the transform of the deviation neglected in the mth integral, the output $X_o(s)$ corresponding to this deviation is

$$X_o(s) = \frac{X_o}{X_i} X_i = \left(\frac{s^m X_o}{X_i} \right) \frac{X_i}{s^m} \quad (65)$$

To evaluate this expression, expand $(s^m X_o/X_i)$ as follows

$$\frac{s^m X_o}{X_i} = a_0 s^m + a_1 s^{m-1} + \ldots + a_m + T_{-(m+1)}(s) \quad (66)$$

To prove that this expansion is correct, compare it with equations 43 and 44. Substitute this expansion into equation 65

$$X_o(s) = a_0 X_i + a_1(X_i/s) + \ldots + a_m(X_i/s^m) + T_{-(m+1)}(s)(X_i/s^m) \quad (67)$$

Take the inverse transform

$$x_o(t) = a_0 x_i(t) + a_1 D^{-1} x_i(t) + \ldots + a_m D^{-m} x_i(t) + \mathcal{R}_{-m}(t) \quad (68)$$

The last term $\mathcal{R}_{-m}(t)$ is the mth integral remainder and represents the convolution of the mth integral curve $D^{-m} x_i(t)$ with the $(m+1)$st integral transient $T_{-(m+1)}(t)$

$$\mathcal{R}_{-m}(t) = T_{-(m+1)}(t) * D^{-m} x_i(t) = \int_0^\infty d\tau T_{-(m+1)}(t) D^{-m} x_i(t - \tau) \quad (69)$$
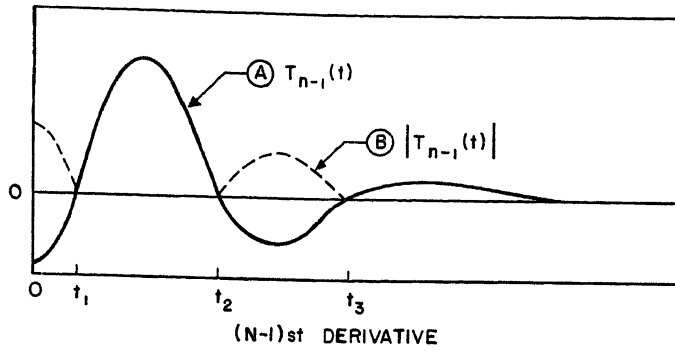
Fig. 13. Method for calculating remainder coefficients

N th DERIVATIVE
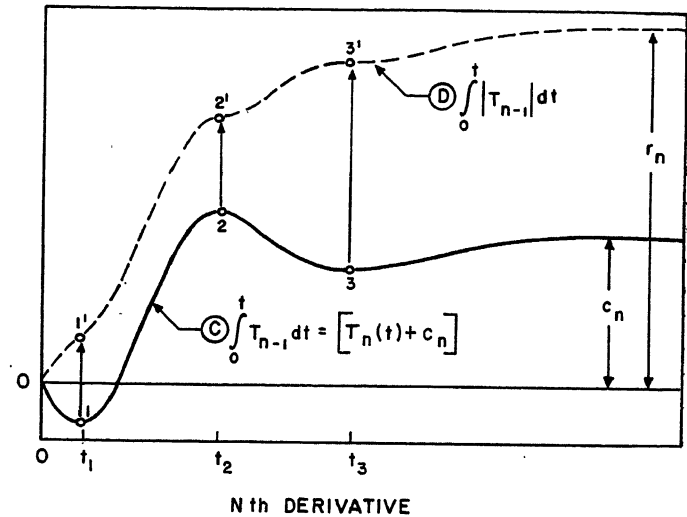
Comparing this with the expression of equation 57 shows that the $\mathcal{R}_{-m}(t)$ is obtained from $\mathcal{R}_n(t)$ by replacing n by $-m$.

When a deviation at a given integral is to be neglected, equation 68 shows that the deviation curve as well as all the lower-order integral curves should be multiplied by the corresponding initial-value coefficient, and the resultant added to the calculated response. The remainder $\mathcal{R}_{-m}(t)$ represents the portion of the output which is neglected. It is not calculated exactly but is bounded by means of the mth-integral remainder coefficient r_{-m}. Since $\mathcal{R}_{-m}(t)$ is obtained by replacing n by $-m$ in $\mathcal{R}_n(t)$, the mth-integral remainder coefficient r_{-m} is obtained by replacing n by $-m$ in the expressions for r_n. Thus, the expression for the remainder coefficient of equation 63 holds for negative values of n as well as for positive values.

CALCULATION OF REMAINDER COEFFICIENT

Method of Calculation

The remainder coefficient r_n can be calculated quite simply by graphical construction from a plot of the corresponding transient term $T_n(t)$. Equation 23 showed that

$$T_n(t) - T_n(0) = \int_0^t T_{n-1}(t)\,dt \tag{23}$$

For $n \geq 0$, the left-hand expression is the composite transient term $[T_n(t)+c_n]$, as shown by equation 24. Compare this with the expression for r_n in equation 63

$$r_n = \int_0^\infty |T_{n-1}(t)|\,dt \tag{63}$$

The composite transient term $[T_n(t)+c_n]$ is the integral of $T_{n-1}(t)$, while the remainder coefficient r_n is the integral of the magnitude of $T_{n-1}(t)$ evaluated at $t = \infty$. Consequently, r_n can be evaluated directly from a plot of $[T_n(t)+c_n]$, as shown by Fig. 13. Curve A

represents the transient term $T_{n-1}(t)$ of a system for the $(n-1)$st derivative, while curve B shows the magnitude of this function. Curve C is the integral of $T_{n-1}(t)$ and is the composite transient term $[T_n(t)+c_n]$, while curve D is the integral of the magnitude of $T_{n-1}(t)$, and its final value is the remainder coefficient r_n.

In the regions where curve C has positive slope, curve D has the same shape as curve C; and in the regions where curve C has negative slope, curve D has the shape of the mirror image of curve C. Thus, to form curve D, the segments of curve C with positive slope and the mirror image of those with negative slope are shifted vertically so that curve D starts at zero and is a continuously increasing curve.

Although this construction was performed upon the transient term for a derivative, the same construction can be performed upon the transient term for an integral. For all cases, the curve for r_n must be started at the origin. It is obvious that the construction can be performed upon either the output-response transient term $T_n(t)$ or the error-response transient term $T_n'(t)$.

Calculation of Remainder Coefficients of Illustrative System

In Fig. 14, the remainder coefficients of the illustrative system are graphically calculated from the transient terms in Figs. 3 to 6, 11, and 12. Fig. 14 shows that r_n is much greater than c_n' for position and velocity, but it is practically equal to c_n' for acceleration, and exactly equal for rate of acceleration and all higher derivatives. The reason that r_n approaches the value of c_n' at large n is apparent from the plots of $T_n'(t)$ for the system which are given in Fig. 15. These plots were obtained by subtracting

the values of the error coefficients from the plots of the composite transient terms. Since the negative area of the position transient term $T_0'(t)$ is roughly equal to the positive area, the sum of the magnitudes of these areas r_1 is obviously much greater than their algebraic sum c_1'. In the velocity transient term $T_1'(t)$, however, the negative area is much less than the positive area; hence, the algebraic sum, the acceleration error coefficient c_2', is only slightly less than the sum of the magnitudes r_2. In the transient terms for acceleration, rate-of-acceleration, and all the higher derivatives, the area is either completely positive or completely negative; therefore, the remainder coefficients for the rate of acceleration and the higher derivatives must be equal to the magnitudes of the error coefficients.

In any adequately damped system, the magnitude of the error coefficient becomes roughly equal to the remainder coefficient at a certain derivative, and often approaches even closer at higher derivatives, but never becomes greater. This first derivative at which c_n is approximately equal to r_n is very significant, because it is at that derivative that a complete straight-line approximation of the input should be made. A basic reason for making the final approximation at that derivative is that the per-unit inaccuracy in the calculated error response is no greater (essentially) than the per-unit approximation of the derivative. For example, in the illustrative system, the final straight-line approximations were made upon the acceleration curve. There is an acceleration component in the computed response equal to c_2' times the approximated acceleration curve, and an acceleration component of inaccuracy for that response equal to r_2 times the acceleration devia-

Fig. 14 (left). Calculation of remainder coefficients of illustrative system

Fig. 15 (above). Transient terms of illustrative system

Fig. 16 (below). Components of deviation for approximations of arbitrary input

nothing is gained by performing this approximation upon a higher derivative.

APPLICATION OF REMAINDER COEFFICIENTS

This section describes the techniques for applying the remainder coefficients. The values of these coefficients for the illustrative systems are

doublet: $r_{-2} = 3.10 \times 10^4$ seconds^{-2} (70)

impulse: $r_{-1} = 182$ seconds^{-1} (71)

position: $r_0 = 1.55$ (72)

velocity: $r_1 = 1.707 \times 10^{-2}$ seconds   (73)

acceleration: $r_2 = 3.57 \times 10^{-4}$ seconds^2 (74)

rate of acceleration: $r_3 = 1.715 \times 10^{-6}$
                 seconds^3 (75)

Fig. 16(C) gives a plot of the acceleration deviation corresponding to the straight-line approximation of the acceleration response. Since the maximum value of the magnitude of the deviation is 55 degrees per second^2, the upper bound to the inaccuracy caused by this approximation is

$$r_2(55° \text{ per second}^2) = (3.57 \times 10^{-4})$$
$$(55)° = 0.0196° \quad (76)$$

It should be noted that the acceleration error-coefficient curve 2 of Fig. 9 used in the calculated response was obtained by multiplying c_2' by the approximated rather than the exact acceleration curve. The reason for this can be seen by examining the remainder expansion of equation 58. The expansion shows that, in order for the remainder coefficient to bound the inaccuracy resulting from neglecting a deviation, that deviation cannot be included in the calculated response. Thus, the inaccuracy bound in equation 76 for the acceleration deviation would not hold if the approximated curve were not employed in the calculated response. Although it might seem that a more accurate result could be obtained by using the exact acceleration curve, this is not so because the acceleration and rate-of-acceleration portions of the response must be derived from the same acceleration curve (i.e., the approximate curve) for the rate-of-acceleration portion to correct properly for the overestimate of the acceleration portion; see Fig. 7.

Step approximations of a derivative should be considered as the integrated effect of line-segment approximations of the lower order derivative. For example, the step approximation of the velocity curve in Fig. 8 at 0.32 second, should be considered as a straight-line approximation of the position curve. In Fig. 17, the solid curve shows the actual input

tion. If the acceleration curve is approximated within 5 per cent, i.e., if the deviation is no greater than 5 per cent of the maximum value of the approximated acceleration curve, the resultant inaccuracy cannot be greater than 5 per cent of the acceleration component of the error response because c_2' is essentially equal to r_2. Since r_2 can never be less than c_2', it is apparent that

Fig. 17 (left). Line-segment approximation of position corresponding to step approximation of velocity

(A) A DEVIATION BOUND FOR nth DERIVATIVE

(B) A TYPICAL REMAINDER TRANSIENT

(C) UPPER BOUND TO INACCURACY FOR (A)

(D) A DEVIATION BOUND FOR nth DERIVATIVE
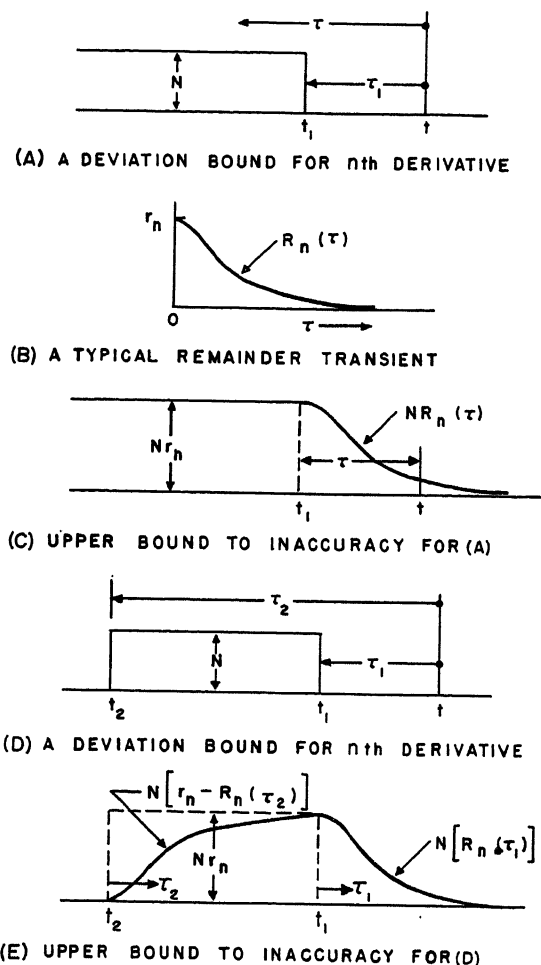
Fig. 19 (right). Application of remainder transient

(E) UPPER BOUND TO INACCURACY FOR (D)

position, and the dashed straight lines show the line-segment approximations which yield the step approximation of velocity when differentiated. The deviation in the position curve corresponding to this approximation is plotted in Fig. 16(A). Similarly, the step approximations of acceleration at 0 and 0.45 second are considered as line-segment approximations of velocity, and the deviations corresponding to these approximations are shown in Fig. 16(B). The deviations in Fig. 16(B) were obtained by integrating the deviations between the approximating steps of acceleration and the actual acceleration curve, just as the deviation of Fig. 16(A) was obtained by integrating the deviation between the approximating step of velocity and the actual velocity curve.

Since the peak magnitude of the veloc-

ity deviation in Fig. 16(B) is 0.57 degree per second, the corresponding inaccuracy bound for this deviation is

$$r_1(0.57° \text{ per second}) = (1.707 \times 10^{-2} \text{ seconds})(0.57° \text{ per second})$$
$$= 0.0097° \quad (77)$$

Similarly, the peak magnitude of the position deviation in Fig. 16(A) is 0.072 degree so that the corresponding inaccuracy bound is

$$r_0(0.072°) = 1.55(0.072°) = 0.112° \quad (78)$$

(A)　(B)　(C)

LINE-SEGMENT APPROXIMATION

PARABOLIC LINE-SEGMENT

LINE-SEGMENT APPROXIMATION OF (N-1)st DERIVATIVE
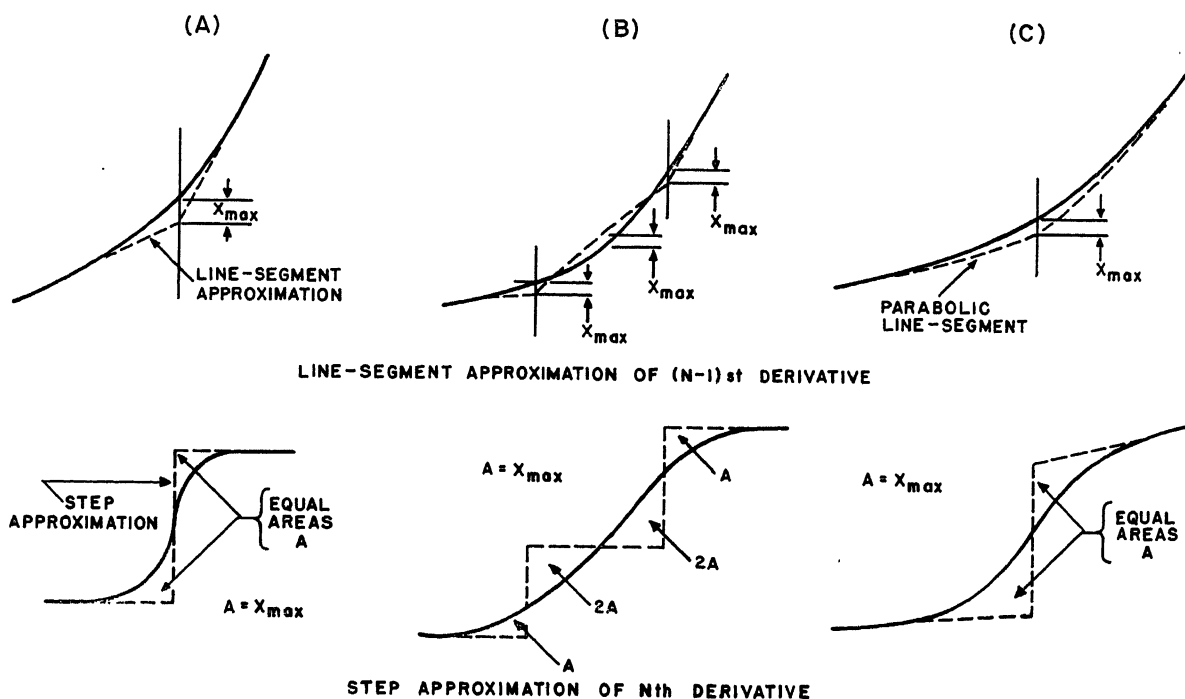
Fig. 18. Comparison of step approximations with equivalent line-segment approximations

STEP APPROXIMATION

STEP APPROXIMATION OF Nth DERIVATIVE

However, these bounds do not quite apply to the calculated error response of Fig. 10 because the line segment approximations were not actually performed; therefore the deviations, rather than being discarded, were included in the calculated response. To correct for this, when step approximations are made, the step deviation should be integrated and the integral multiplied by the corresponding (lower order) output-response coefficient c_n and the product added to the calculated output response. If the error response is being calculated, this product should be subtracted from the error response. In correcting for these deviations, the coefficients for the output response rather than for the error response should be used, because the remainder coefficient is defined in terms of the output response.

Therefore, for the inaccuracy bounds of equations 77 and 78 to apply to the calculated error response of Fig. 10, the calculated response must be corrected by multiplying the deviations of Figs. 16(A) and (B) by the corresponding coefficients c_0 and c_1, and by subtracting them from the error-response curve. The dashed segment 4 in Fig. 10 shows the resultant curve when the position-deviation correction is subtracted. On the other hand, the correction for the velocity deviation of Fig. 16(B) is less than 0.001 degree which is so small with respect to the total error response that it can be neglected. This occurs because the velocity steady-state coefficient c_1 is much smaller in magnitude than the velocity remainder coefficient r_1. Thus, with the corrected segment 4 added to the calculated error response of Fig. 10, the total accuracy bound for the resultant response is the sum of equations 76, 77, and 78

$$0.0196° + 0.0097° + 0.112° = 0.141° \quad (79)$$

This bound applies for all values of time. A more optimistic evaluation of the accuracy can be obtained by applying the remainder transient, described in the next section, which allows a lower inaccuracy bound except for a short period of time near 0.3 second.

When neglecting a deviation, one should always think in terms of the output response, even though the error response is being calculated, because the remainder expression is defined in terms of the output response. In neglecting a deviation of the input position curve one effectively assumes that the output response to that deviation is zero, which requires that the corresponding error response, defined as the input minus the output, must be equal to the deviation itself. Thus, when the error response is

being calculated, all deviations of the input position curve must be aded to the calculated error response.

Although it is convenient to consider a step approximation as a line-segment approximation of the lower order derivative, it is not necessary to perform the line-segment approximation directly. As an example, consider Fig. 18(A) which shows a portion of the $(n-1)$st derivative of a function, with a line-segment approximation and the corresponding portion of the nth derivative with the resultant step approximation. This same step approximation can be obtained directly from the nth derivative as shown by making the areas between the exact curve and the approximating step equal. This area A, is equal to the maximum value of the deviation of the $(n-1)$st derivative from the line-segment approximation X_{max}. The upper bound to the inaccuracy corresponding to this step approximation is obtained by multiplying the area A (deviation X_{max}) by the remainder coefficient of the $(n-1)$st derivative.

If it were necessary to form a closer approximation, two or more steps could be employed instead of one. Fig. 18(B) shows how to approximate a portion of the nth derivative by two steps, and how this approximation appears when performed directly upon the $(n-1)$st derivative. The step approximations are made as before so that the net area of the deviation is zero. However, to obtain the smallest inaccuracy, the second and third areas should be twice the size of the first and last, so that the magnitudes of the three maximum deviations X_{max}, in the equivalent approximation of the $(n-1)$st derivative, are all equal. This maximum deviation X_{max} is equal to the smaller area A of the step approximation.

Strictly speaking, the two step approximations of the acceleration curve in Fig. 8 are not actually steps because the slopes of the approximation preceding and following the discontinuities are not zero, except for the first segment; and, consequently, they correspond to parabolic line-segment approximations of the velocity curve rather than to straight-line approximations. As an illustration, Fig. 18(C) shows a similar step approximation of the nth derivative and the equivalent approximation of the $(n-1)$st derivative. It shows that the $(n-1)$st derivative must be approximated by parabolic line segments to yield, upon differentiation, the step approximation shown upon the nth derivative. Nevertheless, parabolic line segments do not have to be plotted because the approximation may

be performed directly upon the nth derivative as shown.

THE REMAINDER TRANSIENT

A basic limitation to the use of the remainder coefficient is that the upper bound to the inaccuracy calculated from it applies for all values of time. The remainder transient term $R_n(t)$ supplements the remainder coefficient by showing quantitatively the period of time over which the inaccuracy must be considered to apply.

Assume that the nth derivative is approximated in such a manner that the magnitude of the deviation is no greater than N for values of time less than t_1, and is zero for values of time greater than t_1

$$|D^n x_i(t)| \leq N, \text{ for } t < t_1 \quad (80)$$

$$D^n x_i(t) = 0, \text{ for } t > t_1 \quad (81)$$

These bounds on the deviation are shown graphically in Fig. 19(A). Now, the general expression for the remainder associated with any approximation is given by equation 57 as

$$\mathcal{R}_n(t) = \int_0^\infty d\tau\, T_{n-1}(\tau) D^n x_i(t-\tau) \quad (57)$$

As shown in Fig. 19(A), the variable τ represents a time variable measured from time t in the direction of negative time. At values of τ less than τ_1, the nth-derivative deviation $D^n x_i$ is zero; hence it is desirable to split the integration of equation 57 into two regions

$$\mathcal{R}_n(t) = \int_0^{\tau_1} d\tau\, T_{n-1}(\tau) D^n x_i(t-\tau) + \int_{\tau_1}^\infty d\tau\, T_{n-1}(\tau) D^n x_i(t-\tau) \quad (82)$$

The first of the two integrals is zero, because the nth-derivative deviation $D^n x_i$ is zero for that range of τ. In the second integral, the magnitude of the nth-derivative deviation is no greater than N. Therefore

$$|\mathcal{R}_n(t)| \leq N \int_{\tau_1}^\infty d\tau\, |T_{n-1}(t)| \quad (83)$$

Define the integral as the nth-derivative remainder transient $R_n(\tau_1)$

$$R_n(\tau_1) = \int_{\tau_1}^\infty d\tau\, |T_{n-1}(t)| \quad (84)$$

Thus

$$|\mathcal{R}_n(t)| \leq N R_n(\tau_1) \quad (85)$$

A comparison of equation 84 with the expression for r_n in equation 63 shows that, when $\tau_1 = 0$, R_n is equal to r_n

$$R_n(0^+) = r_n \quad (86)$$

Fig. 19(B) shows a plot of a typical remainder transient term $R_n(\tau)$. Its value must be r_n at $\tau = 0$, and its slope can
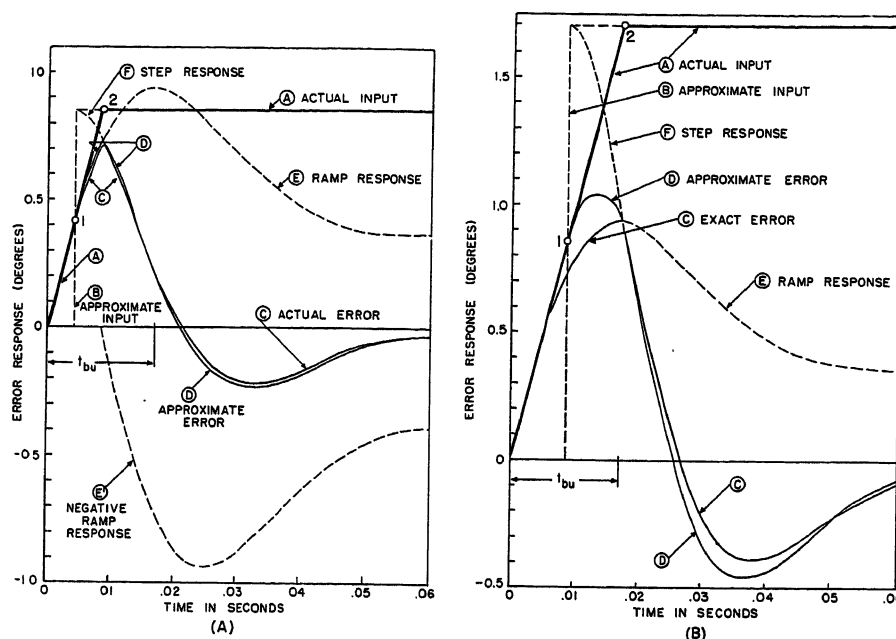
Fig. 20. Comparison of approximate and exact responses to rapidly rising input

never by positive except at $\tau = 0^-$. Fig. 19(C) is a plot of the upper bound to the inaccuracy obtained by applying this remainder transient term to the deviation bound of the nth derivative given in Fig. 19(A). The plot of Fig. 19(C) shows that over the region of approximation the upper bound to the inaccuracy is still Nr_n but, after this region, this upper bound decays following the curve $NR_n(\tau)$.

To calculate the remainder transient term R_n, expand equation 84 as follows

$$R_n(t) = \int_0^\infty dt \,|\, T_{n-1}(t)\,| - \int_0^t dt \,|\, T_{n-1}(t)\,| \tag{87}$$

$$= r_n - \int_0^t dt \,|\, T_{n-1}(t)\,| \tag{88}$$

The variable τ has been replaced by t for convenience. The integral of equation 88 can be obtained from the transient term $T_n(t)$, as was shown in Fig. 13; hence $R_n(t)$ can be obtained by subtracting this integral from its final value, the remainder coefficient r_n. In Fig. 14 the remainder transient terms of the illustrative system are plotted.

The remainder transient shows how the inaccuracy bound decays after the approximation has been ended. It is also desirable to know how the inaccuracy bound builds up at the beginning of the approximation. To determine this, assume that the magnitude of the deviation of the nth derivative is bounded by N for a definite period of time $t_2 < t < t_1$, and is zero elsewhere, as illustrated in Fig. 19(D).

$$D^n x_t(t) = 0, \text{ for } t < t_2, t > t_1 \tag{89}$$

$$|\, D^n x_t(t)\,| \leq N, \text{ for } t_2 < t < t_1 \tag{90}$$

Fig. 19(D) defines the time variables

τ_1 and τ_2 in terms of t_1, t_2, and t. The integral for $\mathcal{R}_n(t)$ in equation 57 may be expressed as the sum of three integrals with the limits

$$\mathcal{R}_n(t) = \int_0^{\tau_1} \cdots + \int_{\tau_1}^{\tau_2} \cdots + \int_{\tau_2}^\infty \cdots \tag{91}$$

Since the expressions within the integrals are the same as in equation 57, they have been omitted for brevity. The first and third integrals are zero because the derivative is zero in these regions of τ, and the second integral yields the following upper bound for $|\,\mathcal{R}_n(t)\,|$

$$|\, \mathcal{R}_n(t)\,| \leq N \int_{\tau_1}^{\tau_2} d\tau \,|\, T_{n-1}(\tau)\,| \tag{92}$$

This bound may be expressed as the difference of two remainder transients

$$|\, \mathcal{R}_n(t)\,| \leq N[R_n(\tau_1) - R_n(\tau_2)] \tag{93}$$

The solid curve in Fig. 19(D) is a plot of the bound of equation 93. This plot shows that at the start of an approximation the inaccuracy bound builds up proportionally to the curve $[r_n - R_n(\tau_2)]$; and at the end of the approximation it decays proportionally to $R_n(\tau_1)$.

Summary of Rules

Approximations in Differentiation Process

Step Approximations

The rule for making step approximations of the input curve or of one of its derivatives is as follows: When a component of the input or of one of its derivatives rises more abruptly than the corresponding transient term for that derivative, that component should be approxi-

mated by a step or if necessary, by a series of steps. In comparing the rate of rise of the two curves a convenient criteria is to compare the times of the two curves to reach 63 per cent for the initial rise (or decay) of the curves.

The choice between approximating a given input by steps, or of differentiating and then approximating, is purely a matter which requires the least labor. In either case, any required accuracy can be obtained by taking sufficiently small approximating segments. The foregoing rule merely points out that, in general, if the derivative curve builds up faster than the transient term, it is easier to approximate by steps; and if it builds up more slowly, it is easier to differentiate before approximating.

The reason for this rule can be explained best by an example. In Fig. 20(A), the heavy curve A represents an input to the illustrative system. Since curve A rises in 0.0085 second, which is one-half the build-up time of the position step response of the system (shown as t_{bu}), the curve is approximated as a step. The dashed curve B is the step approximation to curve A, and is made so that the net area between the two curves is zero, which requires that point 1 be half way between the origin and point 2.

The exact error response to the input A is curve C which is the sum of the positive ramp response curve E, obtained from the velocity transient of Fig. 4, and the displaced negative ramp response curve E'. Curve D is the error response calculated from the step approximation and is the sum of the error response F for the step B, obtained from the position transient of Fig. 3, and the deviation of the actual input A from the approximate input B.

The difference between the exact and approximate responses, curves C and D, has a maximum value of only 0.02 degree, which is only about 3 per cent of the maximum value of the error response. In Fig. 20(B), a similar set of curves is plotted for an input having a build-up time equal to the system build-up time t_{bu}. In this case the deviation between the exact response, curve C, and the approximate response, curve D, has a maximum value of 0.18 degree, which is 19 per cent of the maximum value of the response.

A comparison between Figs. 20(A) and (B) thus shows that the inaccuracy caused by the step approximation is small when the build-up time of the input is somewhat less than t_{bu}, and that it becomes quite large if the input build-up time is significantly greater than t_{bu}.

Biernson—Method for Calculating the Time Response of a System

Final Step Approximation of Derivative

The final step approximation should be made at the lowest derivative, except the first derivative, where the composite transient term $[T_n(t)+c_n]$ is at no time greater in magnitude than the linear term $c_{n-1}t$ resulting from the coefficient c_{n-1} of the lower order derivative. If the rule were expressed in terms of the error response, it would not have to specify that the final step approximations should not be made at the first derivative because, for the first derivative, the lower order error coefficient is c_0', which is generally zero.

Another way to express this rule can be developed as follows. By equation 19

$$c_{n-1} = -T_{n-1}(t)\big|_{t=0+}, \text{ for } n \geq 0 \qquad (94)$$

and by equation 23

$$T_{n-1}(t) = \frac{d}{dt} T_n(t) \qquad (95)$$

Combining these two equations gives

$$c_{n-1} = -\frac{d}{dt} T_n(t)\bigg|_{t=0+}, \text{ for } n \geq 0 \qquad (96)$$

Thus, the error coefficient for the $(n-1)$st derivative is equal to the negative of the initial slope of the nth-derivative transient term. Hence the rule for final step approximations can be expressed as: The final step approximations should be made at the lowest derivative, greater than the first derivative, for which the transient term is at no time greater in magnitude than a straight-line drawn tangent to the transient term at $t=0$.

The final step approximation is achieved by making a complete straight-line approximation of the lower order derivative. For that lower order derivative, the remainder coefficient will be found to be not much greater in magnitude than the steady-state coefficient, provided the lowest frequency set of poles (or pole) is reasonably well damped. Consequently, the per-unit accuracy required in the final straight-line approximation is not much greater than that needed for the calculated error response. For example, for the calculated response to be accurate to within 10 per cent of the maximum error between x_i and x_o, the final straight-line approximation does not have to approximate the derivative by much more than 10 per cent of the maximum value of the derivative curve.

APPROXIMATIONS IN INTEGRATION PROCESS

It has been shown that high-frequency components of the input should be integrated to determine their response.

However, the author has not yet had time to investigate this process in sufficient detail to present definite rules of procedure. Preliminary results of a study of this problem[9] have led to the following tentative rules.

When to Integrate a Component

A component of the input curve or of one of its integrals should be integrated if it rises and falls (like a pulse) in a period significantly less than the time for the corresponding transient term $T_n(t)$ (or composite transient term $[c_0+T_0(t)]$ for the input curve) to reach 63 per cent of its maximum value. Components which rise and fall more slowly should be approximated by steps.

When to Approximate by Steps

Before integrating, step approximations should be made to subtract out any bias components so that the resultant curve to be integrated oscillates about zero with roughly zero net area. If this is not done, the bias components add linearly increasing terms to the higher order integral. A convenient way to accomplish this subtraction is to make a reasonable attempt to remove the constant bias components and to integrate the remaining curve. From the integral, any linearly increasing bias components which are longer in duration than the time for the corresponding transient term to settle to a small per cent of its maximum value are subtracted out. These linearly increasing bias terms are then differentiated and considered as steps of the lower order integral.

Highest Integral Required

If the integration process is performed as described in the foregoing, it appears that the highest integral that need be considered is the next one beyond the lowest integral for which the maximum magnitude of the transient term is its initial value. If the initial value is greater in magnitude than any other point on the transient term, the initial value coefficient a_m is not much smaller in magnitude than the remainder coefficient r_{-m}, provided the highest frequency poles of the system are reasonably well damped.

The reason for stopping at this integral is that, for all components so high in frequency that they are not approximated by steps, when the foregoing procedure is followed, the response is very well characterized at that point by terms due to the initial value coefficients. The inaccuracy resulting from neglecting the higher integrals of these components

is quite small with respect to the initial-value-coefficient terms already included.

CONSTRUCTION OF RESPONSE CURVE

The calculated response is obtained by summing together the following three portions:

1. The transient portion is obtained by multiplying the magnitudes of the various steps of the input, its derivatives, and its integrals by the corresponding transient terms $T_n(t)$.

2. The steady-state portion is obtained by multiplying the approximated input and derivative curves by the corresponding steady-state coefficients c_n.

3. The initial-value portion is obtained by multiplying by the corresponding initial value coefficients a_m the complete integral curves minus the portions included in the step approximations.

It should be emphasized that, when a portion of a derivative curve is not differentiated further, that portion is not included in the steady-state coefficient part of the calculated response, as is shown by the expansion of equation 58. In contrast, when a portion of an integral curve is not integrated further, that portion is included in the initial-value coefficient of the response, as shown by the expansion of equation 68.

DETERMINATION OF ACCURACY OF CALCULATED RESPONSE

If the approximation procedure is performed properly, the inaccuracy in the calculated response is automatically maintained at a low value. A bound on the inaccuracy caused by the deviation at a given derivative or integral is obtained by multiplying the maximum value of the magnitude of that deviation by the corresponding remainder coefficient. The resultant inaccuracy bonds for the input curve and for the various derivatives and integrals are summed to obtain the total inaccuracy bound. Since this bound holds for all values of time, however, a more optimistic bound can be obtained by employing the remainder transient, which shows for how long the corresponding inaccuracy bound applies after a deviation has ended.

TRANSIENT TERMS REQUIRED

It has been shown that for any given system the differentiation process needs to be carried only to a certain derivative and the integration process only to a certain integral, regardless of the input signal. The series of transient terms from that derivative to that integral, therefore, comprise a small finite set of terms which completely define the system time response for any input to essentially
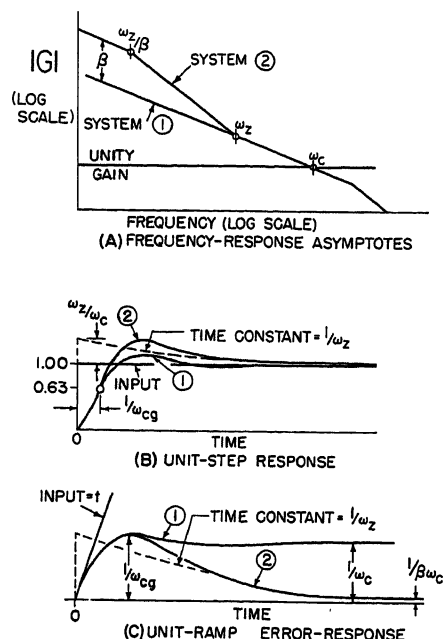
Fig. 21. Relations between transient responses and frequency response

the same degree of (accuracy with respect to either the error response or the output response) as the transient terms themselves are defined.

To illustrate this conclusion, consider an automatic-tracking radar antenna. There are two components in the radar-tracking signal, the actual target signal and the radar noise. Since the tracking bandwidth must be relatively high with respect to the actual target signal for the antenna to follow the target, only the position and derivative transient terms are needed for calculating the response to that input. If these transient terms are specified to only 10 per cent accuracy, the tracking error response to the target signal can still be calculated to essentially 10 per cent of the maximum tracking error. Since the maximum tracking error may be less than 1.0 per cent of the total tracking angle, this can represent less than 0.1-per-cent accuracy with respect to the total tracking angle. In like manner, the tracking bandwidth must be low with respect to the radar noise so that the noise does not excessively disturb the tracking line; hence only the position and integral transient terms are required for calculating the response to the radar noise. If these transient terms are specified to within 10 per cent, the error in tracking produced by the noise can be calculated essentially to within 10 per cent of the maximum tracking error. This can represent a much higher percentage accuracy in terms of the input noise signal, because the tracking re-

sponse to the noise is considerably smaller than the input noise signal.

This example illustrates that when the complete set of transient terms are given it is not necessary to specify them very accurately. Relatively crude approximations of the transient terms are all that is required to yield adequate engineering information concerning the system time response to any input signal.

Theoretically, of course, the time response of a system is completely specified by the impulse response, or by the step response, or by any transient term of the system because, if any one term is completely specified, the others can be calculated by differentiation and integration. However, when only one transient term is given, it must be specified to a very high degree of accuracy to yield sufficient accuracy in the other terms.

Relationships Between Transient Terms and Loop Frequency Response

The application of the approach of this paper to the design of feedback-control systems is greatly enhanced by a firm understanding of the relationships between the transient terms of a system and its loop frequency response. The major characteristics of the transient terms can be approximated quite readily merely by inspecting a plot of the loop frequency response. A number of these approximations have been described in detail by the author,[1] and the most important of these are summarized in the following. They are expressed in terms of the basic variables of feedback-control loops defined in Fig. 1(B). A knowledge of the relationships between the closed-loop poles and the open-loop transfer function given in reference 6 greatly aids in justifying these approximations.

BANDWIDTH

The gain-crossover frequency of a feedback-control loop ω_{cg} is defined as the frequency where the magnitude of the loop frequency response $|G(j\omega)|$, called the loop gain, crosses the unity gain axis. It is the best criteria for the bandwidth of the loop because it specifies very reliably the following characteristics of the system transient terms:

1. The time for the step response to rise to 63 per cent of its final value (the rise time) is roughly $1/\omega_{cg}$.

2. The maximum error for a unit ramp (step of velocity) is roughly $1/\omega_{cg}$.

3. The maximum response to a unit impulse is roughly ω_{cg}.

Approximations 1 and 2 hold quite

well for all reasonably stable feedback-control loops. Approximation 3 also holds provided the loop gain attenuates reasonably fast above gain crossover, which is quite desirable for noise considerations.

LOW-FREQUENCY INTEGRATIONS

The effect of a low-frequency integral network is illustrated in Fig. 21. Systems 1 and 2 have the same asymptote-crossover frequency ω_c; hence they have essentially the same bandwidth because the actual gain-crossover frequencies ω_{cg} are nearly equal to ω_c. At zero frequency, both systems have only a single integration, but system 2 has a double integration for a significant frequency region below ω_c, which produces a much higher velocity constant (lower c_1') than for system 1.

The approximate error responses to a unit ramp (composite error-response transient terms) for the two systems are shown in Fig. 21(C). Since the crossover frequencies are the same, the maximum values of error are about the same and are approximately equal to $1/\omega_{cg}$. Since the velocity constant of system 1 is ω_c, the unit-ramp response for system 1 approaches a final value c_1' equal to $1/\omega_c$. However, the double integration in system 2, by increasing the velocity constant to $\beta\omega_c$, reduces the final value of its ramp response to $1/\beta\omega_c$. Since the speed of the integrating action of the second low-frequency integration in system 2 is determined by its upper break frequency ω_z, which defines a low-frequency zero of G, the tail in the ramp response for system 2 approximately follows an exponential with the time constant $1/\omega_z$.

The acceleration error-response transient for system 2 is approximately a rising exponential with a time constant of $1/\omega_z$ and a final value c_2' of roughly (c_1'/ω_z), which is equal to $(1/\beta\omega_c\omega_z)$.

The step responses of the two systems are shown in Fig. 21(B). Since both systems have the same crossover frequencies, the responses rise to 63 per cent of the final values in essentially the same time $1/\omega_{cg}$. The main difference between the two responses is that system 2 has an additional exponential component which increases the overshoot and produces a tail. The magnitude of this exponential component is roughly (ω_z/ω_c), and its time constant is roughly $1/\omega_z$.

STOPPING POINTS IN DIFFERENTIATION AND INTEGRATION

The stopping points in the differentiation and integration process can be ex-

pressed quite simply in terms of the asymptotes of the loop transfer function G as follows:

1. The order of derivative at which the final straight-line approximations are made is equal to the maximum number of integrations of the loop transfer function at frequencies below gain crossover; these integrations do not necessarily have to apply at zero frequency.

2. The order of integral at which the integration process should be stopped is one greater than the maximum number of integrations of the loop transfer function at frequencies above gain crossover.

Conclusions

COMPARISON WITH OTHER METHODS

What is the basic difference between the method presented and other methods for calculating the response of a system to an arbitrary input? Essentially all of them break the input down into a series of steps of position, integrals of position, and/or derivatives of position. However, this method differs basically from the others in that it employs an optimum method of breaking the input down. Consequently, for a given degree of accuracy in the calculated response, far less segments need be employed and far less accuracy is needed in most of the calculations. The resultant analytical technique is therefore much quicker and easier to apply.

The most important advantage of this method over previous methods is that it yields a direct means of synthesizing a feedback-control system for a prescribed response to a particular input. Since the response calculated is constructed from a relatively small number of portions, which are related to the derivatives of the input by means of the system transient terms, the designer can usually tell what transient terms should be modified to improve the response. Approximations such as those presented in the previous section allow the designer to relate the forms of the transient terms directly to the open-loop frequency response. Thus, with this method it is fairly simple to determine the required open-loop frequency response for the system to yield a prescribed response to a particular arbitrary input.

An important aspect of this method is the expression it gives for the inaccuracy bound in the calculated response. Since this inaccuracy bound is quite easily calculated by means of the remainder transient or remainder coefficient, one can easily determine at any stage of the calculation how accurately the differentiation should be performed, and how

closely the approximations need to be made. A rough estimate of the response is frequently all that is necessary, and so it is very important that approximate calculations can be made with the certainty that the resultant calculated response has a sufficient degree of accuracy. Besides, the remainder transient has another important use when there is inaccuracy in the input data, because it shows what the corresponding inaccuracy is in the system response.

GENERAL SPECIFICATION OF TIME RESPONSE OF SYSTEMS

In the past it has been very difficult to specify the time response of a system in a general manner. The most common means of specification have been the rise time, peak overshoot, and settling time of the response of a system to a step of position. The settling time and peak overshoot of the ramp response and some of the error coefficients, particularly the velocity and acceleration coefficients, have also been found quite useful. Even though these parameters describe certain important aspects of the system time response, there has up to now been no adequate means for describing the time response completely in a general fashion.

To be sure, the time response of a system to any input theoretically can be calculated from the impulse response, or from the step response, or from any transient term of the system. However, this requires that the particular transient term be specified to a very high degree of accuracy, far greater than the accuracy needed in the calculated response. Consequently, it is impossible to set rough bounds upon the major characteristics of the particular transient term and define the time response of the system in a general fashion. For example, it would be very difficult to restrict the shape of the impulse response by a few parameters in such a manner as to maintain a particular velocity constant for the system.

This paper shows a general method for specifying the time response of a system. The time response of a system is specified by the complete set of transient terms. If the complete set is known, each term need be specified only to the same degree of accuracy as is needed in the calculated response. Since there are in practice only a few terms in this set, it is perfectly feasible to specify them all.

RELATIONSHIP BETWEEN FREQUENCY RESPONSE AND RESPONSE TO ARBITRARY INPUT

Since the transient terms are a practical means of specifying the system time

response, they provide an excellent media for relating the frequency response of a given system to its general time response. This is particularly true because adequate approximations of the transient terms can generally be obtained by inspecting the open-loop frequency response.

For example, the system gain-crossover frequency determines the peak value of the impulse response, the rise time of the step response, and the peak error in the ramp response. An additional low-frequency integration increases the peak overshoot in the step response and reduces the final value of the error in the ramp response. The upper break frequency of this integration determines the settling time of a tail in the step response, the settling time of the ramp error response, and the build-up time and final value of the error response to a step of acceleration. In a similar fashion, it is possible to relate all the important characteristics of the system open-loop and closed-loop frequency responses to the forms of the system transient terms.

Thus, this method shows directly what it means in terms of the time response of a system to modify the system frequency response in a particular manner. For example, the error which results when a system follows a given input can be decreased either by increasing the bandwidth of the system or by increasing the low-frequency gain. This condition has often led to the belief that one method is as effective as the other in reducing the system errors, but this in general is not so. Additional low frequency gain reduces constant velocity components of error by integrating them out after a period of time whereas additional bandwidth reduces velocity components of error by maintaining them small from the start. These effects can be deduced from the velocity transient term. The bandwidth determines the maximum value of the error in the velocity transient term, and the low-frequency portion of the system transfer function determines the final value of the error and the settling time.

SIGNIFICANCE OF ERROR COEFFICIENTS

Many attempts have been made to characterize completely the time response of a system to an arbitrary input by means of error coefficients. Some of these attempts have been quite successful and others unsuccessful. One fundamental difficulty is that the error coefficient term for a given derivative cannot strictly characterize a derivative which has a discontinuity. Consequently, it has been

assumed in practice that following a discontinuity in a given derivative, the error coefficient term applies only at values of time later than the system settling time. However, the actual value of the system settling time has not been defined explicitly and it has been incorrectly assumed that it represents the time for the position step response to reach a given percentage of its final value.

The error coefficient expansion has been derived as an infinite series which might lead to the assumption that the response to any input can be computed merely by employing enough terms in the expansion, but this is not so. In fact, experience has shown that only a very few error coefficients are significant.

The implications of the error coefficient approach are explained quite simply by the method of this paper. The error coefficient expansion should be considered as merely a portion of the total response, the other portion being the sum of the transient terms. It is obvious, then, that the error coefficient portion is an adequate representation of the response only when the transient terms are relatively small. If there is a discontinuity in a particular derivative, the transient term for that derivative must settle out before the error coefficient term can be applied by itself. Besides, since there is only a finite set of transient terms which are significant, only a finite set of error coefficients are significant.

The concept of a system settling time is erroneous. Rather, there is a transient settling time for each derivative. If there is a step of position one must wait until the position transient has settled out before applying the position coefficient; and if there is a step of acceleration one has to wait until the acceleration transient has settled before applying the acceleration coefficient. For example, the step response of Fig. 3 remains within 5 per cent of its final value after 0.05 second, while the time for the step-of-acceleration transient of Fig. 5 to reach 5 per cent of its final value is 0.15 second. If the value ω_2 in the system transfer function, Fig. 2, were set much lower with respect to ω_c, the discrepancy between these two settling times would be much greater.

Actually, the error coefficients should be considered as no more or no less than the final values of the corresponding transient terms. A system should not be designed to have a given error coefficient but rather to have a particular transient term. For instance, the velocity constant of a system should not be specified by itself. Instead, the maxi-

mum value of the velocity transient (roughly equal to $1/\omega_{cg}$), the settling time of the velocity transient, and, its final value, the velocity error coefficient should be specified. It is obvious that two systems with the same bandwidth and the same velocity constant will generally behave quite differently if the velocity transient of one of them has a much longer settling time than that of the other.

Appendix I. Computation of Steady-State and Initial-Value Coefficients by Long-Division Expansion

The long division process for calculating the steady-state coefficients c_n and the initial-value coefficients a_m is best described by an example. Given the transfer function

$$\frac{X_o}{X_i} = \frac{5+5.5s}{(s+1)(s+5)} = \frac{5+5.5s}{5+6s+s^2} \qquad (97)$$

The steady-state coefficients are obtained by dividing the numerator of X_o/X_i by the denominator in a long division process as follows

$$
\begin{array}{r}
1-0.1s-0.08s^2+0.116s^3 \\
5+6s+s^2 \overline{)5+5.5s} \\
\underline{5+6s+s^2} \\
-0.5s-s^2 \\
\underline{-0.5s-0.6s^2-0.1s^3} \\
-0.4s^2+0.1s^3
\end{array} \qquad (98)
$$

etc. Thus X_o/X_i is expanded to the series

$$\frac{X_o}{X_i} = 1-0.1s-0.08s^2+0.116s^3\ldots \qquad (99)$$

so that the steady-state coefficients must be

$$c_0 = 1 \qquad (100)$$

$$c_1 = -0.1 \text{ second} \qquad (101)$$

$$c_2 = -0.08 \text{ second}^2 \qquad (102)$$

$$c_3 = +0.116 \text{ second}^3 \qquad (103)$$

If the order of the polynomials is reversed in the long division so that the numerator is $(5.5s+5)$ and the denominator is (s^2+6s+5), the long division yields the expansion

$$\frac{X_o}{X_i} = \frac{5.5}{s}-\frac{28}{s^2}+\frac{140.5}{s^3}-\ldots \qquad (104)$$

Hence, the initial-value coefficients are

$$a_0 = 0 \qquad (105)$$

$$a_1 = 5.5 \text{ second}^{-1} \qquad (106)$$

$$a_2 = -28 \text{ second}^{-2} \qquad (107)$$

$$a_3 = 140.5 \text{ second}^{-3} \qquad (108)$$

Appendix II. Calculation of Transient Terms for Illustrative System

The transfer function X_e/X_i for the illustrative system is calculated from the loop transfer function G in equation 32 as follows

$$\frac{X_e}{X_i} = \frac{1}{1+G}$$

$$= \frac{s(s+0.0307\omega_c)(s+1.23\omega_c)}{1.23\omega_c^2(s+0.153\omega_c)+s(s+0.0307\omega_c)} \qquad (109)$$

By the methods of reference 6 the denominator can be factored to the form shown in equation 34. The error response $X_e(s)$ for a unit step input is thus

$$X_e(s) = \frac{1}{s}\frac{X_e}{X_i}$$

$$= \frac{(s+0.0307\omega_c)(s+1.23\omega_c)}{(s+0.1750\omega_c)(s^2+2\zeta\omega_n s+\omega_n^2)} \qquad (110)$$

Equation 110 may be expanded in terms of the poles as follows

$$X_e(s) = \frac{K_{10}'}{s+0.1750\omega_c}+\frac{K_{20}'}{s+\zeta\omega_n-j\omega_0}+\frac{K_{20}'^*}{s+\zeta\omega_n+j\omega_0} \qquad (111)$$

where

$$\omega_0 = \sqrt{1-\zeta^2}\,\omega_n = 0.880\omega_c = 108.8 \text{ radians per second} \qquad (112)$$

Take the inverse transform of $X_e(s)$ in equation 111

$$x_e(t) = K_{10}'\epsilon^{-0.175\omega_c t}+\left|2K_{20}'\right|\epsilon^{-\zeta\omega_n t}\times \sin(\omega_0 t+\varphi_{20}') \qquad (113)$$

where the angle φ_{2n}' is defined as

$$\varphi_{2n}' = \angle K_{2n}'+90° \qquad (114)$$

Equation 113 is $[c_0'+T_0'(t)]$, which is equal to $T_0'(t)$ since c_0 is zero. The other transient terms have the form

$$T_n'(t) = K_{1n}'e^{-0.175\omega_c t}+\left|2K_{2n}'\right|\epsilon^{-\zeta\omega_n t}\times \sin(\omega_0 t+\varphi_{2n}') \qquad (115)$$

Table I. Coefficients for Illustrative System

n	K_{1n}'	$\|2K_{2n}'\|$	φ_{2n}', Degrees
-2	-76.8 second^{-2}	4.36×10^4 second^{-2}	-57.3
-1	3.57 second^{-1}	343 second^{-1}	-178.8
0	-0.166	2.70	59.7
1	7.71×10^{-3} second	2.120×10^{-1} second	-61 8
2	-3.58×10^{-4} second^2	1.67×10^{-4} second^2	176.7
3	1.66×10^{-5} second^3	1.32×10^{-6} second^3	55 2

Employ equation 17 to relate K_{1n}' and K_{2n}' to K_{10}' and K_{20}'

$$K_{1n}' = \frac{K_{10}'}{(-0.175\omega_c)^n} \tag{116}$$

$$K_{2n}' = \frac{K_{20}'}{(-\zeta\omega_n + j\omega_0)^n} = \frac{|K_{20}'| \angle \varphi_{20}' - 90°}{\omega_n{}^n / n(180° - \cos^{-1}\zeta)} \tag{117}$$

By equation 117, φ_{2n}' is related to φ_{20}' by

$$\varphi_{2n}' = \varphi_{20}' + n(\cos^{-1}\zeta - 180°) \tag{118}$$

The coefficients K_{10}' and K_{20}' are calculated from the transfer function X_e/X_t in equation 34 by an equation equivalent to equation 16, and they yield the values for K_{10}', $|2K_{20}'|$, and φ_{20}' given in Table I for

$n = 0$. The values for other values of n given in Table I can be obtained from these by applying equations 116, 117, and 118.

References

1. A Simple Method for Calculating the Time Response of a System to an Arbitrary Input. G. Biernson. *Report No. 7138-R-3*, Massachusetts Institute of Technology, Cambridge, Mass., Jan. 20, 1954, ASTIA AD-25082.

2. Transients in Linear Systems, Vol. I (book), M. F. Gardner, J. L. Barnes. John Wiley and Sons, Inc., New York, N. Y., 1942, pp. 155.

3. Theory of Servomechanisms (book), H. M. James, N. B. Nichols, R. S. Phillips. McGraw-Hill Book Company, Inc., New York, N. Y., 1947, section 4-4.

4. Principles of Servomechanisms (book), G. S. Brown, D. P. Campbell. John Wiley and Sons, Inc., New York, N. Y., 1948, p. 167.

5. The Decilog: A Unit for Logarithmic Measurement, E. I. Green. *Electrical Engineering*, vol. 73, no. 7, July 1954, pp. 597–99.

6. Quick Methods for Evaluating the Closed-Loop Poles of Feedback Control Systems, G. Biernson. *AIEE Transactions*, vol. 72, pt. II, May 1953, pp. 53–70.

7. A Note on the Error Coefficients of a Servomechanism, J. L. Bower. *Journal of Applied Physics*, New York, N. Y., July 1950.

8. A Closed Expansion of the Convolution Integral (A Generalization of Servomechanism Error Coefficients), E. Arthurs, L. H. Martin. *Ibid.*, Jan., 1955, pp. 58–60.

9. A Graphical Solution of the Response of a Feedback Control System to an Arbitrary Input. George J. Mealey. *M.S. Thesis*, Massachusetts Institute of Technology, Cambridge, Mass., June 1955.

Relay Response to Motor Residual Voltage During Automatic Transfers

A. R. KELLY
ASSOCIATE MEMBER AIEE

Synopsis: Undervoltage relays and starter solenoids respond to motor residual voltage. Test data are presented showing how such response can affect operation of an automatic transfer in a secondary selective system supplying induction motors. Suggestions are given for relay application.

INDUSTRIAL applications are found for power systems which automatically transfer motor loads from a normal to an alternate supply upon loss of the normal supply. If the motors drive fluid-handling equipment, it may be desirable to restore power very rapidly for two reasons: First, a shorter outage reduces loss of energy in liquid velocities and gas pressures; this decreases the magnitude and shortens the time of the reaccelerating energy demand on the alternate supply after transfer, thus decreasing system disturbance. Second, if the fluids are being heated, cooled, or chemically processed, a shorter outage can reduce severity of equipment overheating and thermal shock, and the extent of undesirable or dangerous chemical reactions. Thus systems are frequently designed to complete the automatic transfer and to begin reaccelerating the motors before the motors can come to rest.

When designing such systems, motor residual voltage should be considered. In this paper the nature of residual voltage, the undesirable subtransient cur-

rents it can cause, and some methods of preventing these currents are reviewed briefly. For the one of these methods believed most applicable to industrial power systems, test data are presented to illustrate certain requirements in the selection and setting of relays associated with the automatic transfer.

Nature of Residual Voltage

If the circuit supplying a running induction motor is opened under load, energy is stored in the magnetic field, mainly in the air gap, and in the rotating mass. This results in a residual voltage at the stator terminals and an ability to deliver electric power to an externally connected load, as in regenerative braking. Magnitude and frequency of residual voltage decrease with time, from initial values slightly less than system voltage and frequency, at rates dependent on several factors.

Those affecting the frequency of residual voltage are:

1. With no electric load, frequency will correspond to shaft speed which drops because of mechanical load, friction, and windage.

2. If the stator supplies an external electric load, the load current in the stator will produce braking torque which adds to the braking torques in factor 1. In addition, if the load is not capacitive, stator load current produces armature reaction which continuously shifts the

magnetic poles in the rotor opposite to direction of rotation, reducing frequency to less than that corresponding to shaft speed.

Factors affecting the magnitude of residual voltage are:

1. With no electric load, voltage will be proportional to the product of two factors: shaft speed, which decreases with time as discussed in the foregoing, and an exponential decrease in rotor flux (of the form $\epsilon^{-Rt/L}$) caused by rotor I^2R and core losses.

2. If an external electric load is supplied, voltage is further reduced by stator-current impedance drop and armature reaction.

These factors affecting residual voltage can be generalized and regrouped under the following headings:

1. Motor characteristics: rotational inertia of motor, friction and windage of motor, and motor time constant. (The motor time constant referred to herein is the open-circuit time constant, the time for voltage to drop to 0.368 of the air-gap voltage with the shaft speed held normal and no electric load.)

2. Other factors: rotational inertia of driven load, torque requirements of driven load, and external electric load, if any.

Often several motors, each with a different time constant, mechanical load, and inertia, are connected to the same bus. If power to that bus is interrupted under load, the motors with shorter time constants and faster drops in shaft speed will continue to "motor," becoming

(A) FULL-VOLTAGE START

(B) REACCELERATION UPON TRANSFER

electric loads for and being fed by the other motors as long as they all remain connected through the bus. There will be one bus voltage and one frequency, both decaying faster than the open-circuit voltage and frequency of the "generators" but slower than that of the "motors." (Capacitors or over-excited synchronous motors connected to the bus may supply excitation current for the induction motors so that they sustain higher residual voltages than otherwise.)

It will be recognized that residual voltage is the cause of "induction-motor contribution to fault current" used in short-circuit calculations. If not separated from the motors by appreciable impedance, a fault constitutes a severe load which dissipates the stored magnetic energy of the motors in a very few cycles. Thus, in automatic transfers occasioned by short circuits, residual voltage usually disappears before transfer is initiated. This paper is concerned with the less frequent transfers not associated with short circuits, wherein supply voltage is suddenly lost such as through switching error or control-circuit failure. In such cases residual voltage may persist several seconds.

Subtransient Reaccelerating Current

In an automatic transfer, an inrush or subtransient reaccelerating current flows when motors are connected to the alternate source. The value of this inrush is approximately proportional to the "resultant voltage." Resultant voltage varies with the magnitude of residual voltage, and with the phase angle between residual and source voltage, as shown in Fig. 1. The symmetrical re-

accelerating current, illustrated in Fig. 1(B), will be modified by offset or initial asymmetry in the line current, as is the case for normal symmetrical starting current illustrated in Fig. 1(A). If the angle between residual and source voltages is less than 60 degrees, the resultant voltage must be less than the source voltage; if more than 90 degrees, the resultant voltage must be greater than the source voltage; if between 60 and 90 degrees, the relation depends on relative magnitudes. If the transfer is not synchronized, probabilities favor a resultant voltage greater than the source voltage, with larger residual voltages causing larger resultant voltages. Thus, assuming subtransient reaccelerating current proportional to resultant voltage, higher residual voltages tend to cause higher subtransient currents. Hence, while faster transfers reduce the over-all reaccelerating demand after transfer, as discussed in the foregoing, they may increase the subtransient current at transfer.

Such subtransient currents are of concern for two reasons. If they are high enough, damage to motor windings and excessive torque may result. Second, if the subtransient reaccelerating current for a motor is higher than the sub-transient starting current, instantaneous overcurrent protection for that motor must be set higher than required for starting. Since motor time-overcurrent protection is usually of the long-time type to permit starting, instantaneous overcurrent setting for the largest motor on a bus often fixes minimum pickup for the short-time overcurrent protection of the feeder to the bus. Thus, increasing the motor instantaneous settings may lead to relay co-ordination problems.

Methods of Reducing Subtransient Reaccelerating Current

The subtransient reaccelerating current can be reduced by reducing the resultant voltage. One way would be to synchronize the transfer so that the residual voltage is within ±60 degrees of the alternate-source voltage at the instant of transfer. The fluctuation in the bus load, hence frequency decay rate, found in most industrial systems would require extremely complex relaying to provide this synchronizing, if indeed it were possible. The alternative is to assure a low magnitude of residual voltage at transfer. A short circuit could be switched on and off the bus before connection to the alternate source, but this would require expensive equipment. Motors can be limited to high-slip designs to reduce time constant, also giving higher reaccelerating torque and lower starting current. .

A simpler method of limiting the reaccelerating current to safe levels is to delay the transfer until the residual voltage drops to a specified value. This value can be established on the basis of Fig. 1(B) as follows: Induction motors are guaranteed for operation at 110 per cent of the motor rated voltage. The reaccelerating current may therefore be at least as high as that which will occur when the voltage at the motor IZ_m is 110 per cent. This will be 105 per cent of the system nominal, or no-load, voltage for systems above 600 volts. This sets 105 per cent as the maximum desirable value of IZ_m in Fig. 1(B). The limiting condition is with the residual voltage 180 degrees from the alternate-source voltage. Using the approximation that IZ_m and IZ_s add algebraically to give the resultant voltage, the maximum desirable value of the residual voltage can be derived from the two summations which equal the resultant voltage

$$\text{resultant} = 100 + \text{residual} = IZ_m + IZ_s \quad (1)$$

$$\text{maximum residual} = 105 + IZ_s - 100$$
$$= IZ_s + 5 \quad (2)$$

with all voltages in per cent of nominal. From equation 2 it is seen that the maximum desirable residual voltage is a function of the system voltage drop on reacceleration. If only one small motor is to reaccelerate, the residual voltage should theoretically be limited to 5 per cent. In most cases, however, the reaccelerating load will cause a considerable voltage drop IZ_s through the system, permitting transfer at higher residual voltages. The ratio $Z_m/(Z_m + Z_s)$, a measure of the amount of load

reaccelerated, will usually fall between 0.65 and 0.9. (Note that Z_m is approximately the locked-rotor impedance of all motors in parallel.) For the ratio 0.65

$$Z_m = 0.65 \, (Z_m + Z_s) \tag{3}$$

$$Z_s = 0.54 \, Z_m \tag{4}$$

$$IZ_s = 0.54 \, IZ_m \tag{5}$$

Substituting equation 5 in equation 1

$$100 + \text{residual} = IZ_m + 0.54 \, IZ_m = 1.54 \, IZ_m \tag{6}$$

$$\text{maximum residual} = (1.54)(105) - 100 = 62 \text{ per cent} \tag{7}$$

For the ratio $Z_m/(Z_m + Z_s) = 0.9$, similar calculations yield a maximum desirable residual voltage of 17 per cent. A value of 25 per cent is recommended for the usual case. This corresponds to an impedance ratio of 0.84 and has been used successfully without trouble from breaker tripping. While some risk is taken for transfers with few motors reaccelerating, transfer times can become quite long if delayed until the residual voltage drops to much below 25 per cent.

Delay in transfer, until the residual voltage reaches 25 per cent, can be achieved by means of fixed time delay; this requires either transfer tests or calculations[1] to determine the operating condition with the most gradual decay of voltage. In addition, a transfer occasioned by a short circuit on the supply system would be needlessly delayed.

The method recommended for industrial power systems consists of sensing the residual voltage by means of a relay designed for this service, which holds up transfer until the voltage drops to 25 per cent. This scheme requires time-delay undervoltage protection for motors with solenoid-type starters if they are to reaccelerate automatically, since 25 per cent is well below drop-out voltage of such starters.

Description of Tests

The power system shown in Fig. 2 was designed for automatic transfer and reacceleration of all motors upon loss of either 13.2-kv supply. The relays shown are described in Table I. If source 1 loses voltage, e.g., circuit breakers C and F trip when relays 1 and 6 time out. Then, when the residual voltage drops sufficiently, relay 4 drops out and permits breaker D to close; similarly when relay 8 drops out, breaker G closes. If the voltage is restored before relays TD drop out, the 480-volt magnetic contactors are re-energized;

Fig. 2. Refinery process unit power system. All 440-volt motors have 3-wire control with time-delay undervoltage protection; see Table I for relay descriptions
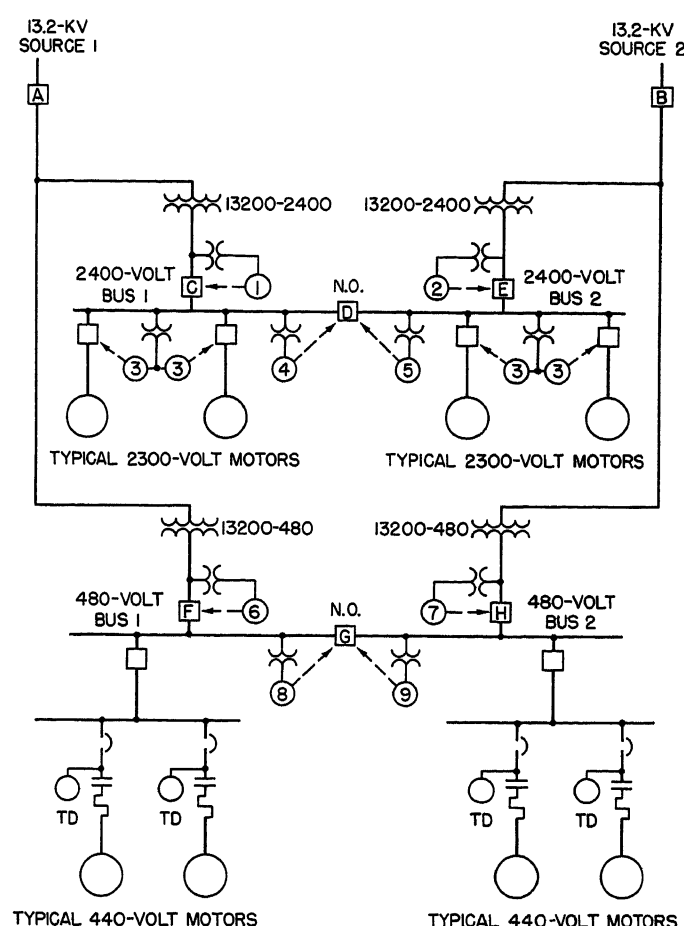
if the voltage is restored before relays 3 time out (they are set for lower voltage and longer time than relays 1 and 2), the 2,300-volt motor breakers remain closed. A 13.2-kv fault below breaker A trips breakers A, C, and F together, initiating transfers without waiting for relays 1 and 6 to time out, since it is certain that the voltage loss is not momentary. On a 2,400- or 480-volt bus fault it is necessary for breakers C, E, F, or H to trip on overcurrent rather than on undervoltage so as to lock out transfer; this co-ordination requirement set the minimum times for relays 1, 2, 6, and 7, which were different for 2,400 and 480 volts, as shown in Table I.

Nine test transfers, designated tests A through J (omitting I) were effected by tripping breaker A or B. In all

but test B the simultaneous-tripping circuits were disconnected so that opening breaker A or B simulated loss of source voltage and transfers were initiated by relays 1 and 6 or 2 and 7. Oscillograms were taken of the voltage on either the 2,400- or 480-volt de-energized bus. A 60-cycle reference trace, an oscillograph event marker, and a cycle counter were used to determine times at which the transformer secondary breaker opened and at which relays 4, 5, 8, or 9 dropped out. Tables II and III show at which bus such measurements were made, and those motors running at each test. All motors drove centrifugal pumps except the five 60-horsepower motors which drove cooling tower fans, and all were under normal load except as noted in the tables. No capacitors

Table I. Description of Undervoltage Relays Shown in Fig. 2

Relay Type	Relay Number	Drop-out Voltage at 60 Cycles	Drop-out Time at 0 Volts, Seconds
Induction disk General Electric type IAV	1, 2	93	1.2
	3	82	1.9
	6, 7	93	1.6
Instantaneous General Electric type HGA	4, 5	26	0
	8, 9	32	0
Durakool type BFT*	TD	215 to 255	1.8 to 2.2

* Mercury bottle relay with restriction orifice, instantaneous pickup, and time-delay dropout, connected directly to 480-volt lines. Values show manufacturing tolerances.

Table II. Data on Tests Opening Circuit Breaker A

Test Number	A	B	C	D	E
Oscillograph location					
2,400-volt bus 1				X	X
480-volt bus 1		X	X	X	
2,300-volt motors					
450 horsepower, 3,600 rpm.	2,400-volt bus 1 fed through breaker D from source 2	R	R	R	N
125 horsepower, 3,600 rpm*		R	R	R	O
125 horsepower, 3,600 rpm.		R	R	R	O
100 horsepower, 3,600 rpm*		R	R	R	N
100 horsepower, 1,800 rpm*		R	R	R	N
440-volt motors					
60 horsepower, 1,800 rpm	R	R	O	O	480-volt bus 1 fed through breaker G from source 2
60 horsepower, 1,800 rpm	R	R	O	O	
50 horsepower, 3,600 rpm	R	R	R	R	
50 horsepower, 3,600 rpm	R	R	O	R	
50 horsepower, 1,800 rpm	R	R	R	R	
40 horsepower, 3,600 rpm	R	R	O	O	
15 horsepower, 3,600 rpm	R	R	R	R	
3 horsepower, 1,800 rpm	R	R	O	R	
2 horsepower, 900 rpm	R	R	R	N	
1 horsepower, 900 rpm	R	R	O	N	

R—Reaccelerated
O—Failed to reaccelerate.
N—Not running during this test.
* Running with discharge nearly closed; spare, running by steam or from other bus, carrying most of load. These footnotes also apply to Table III.

or synchronous motors were connected.

Figs. 3, 4, and 5 show voltage and frequency data obtained from the oscillograms. Voltage magnitudes were calculated from peak-to-peak amplitudes, neglecting any asymmetrical offset. Values below about 20 volts could not be read accurately, so the curves below this are extrapolated. Frequency was calculated, for several points in time, from the average frequency 2 to 5 cycles before and after each point. Estimated accuracy of the curves is ±2 cycles for frequency, ±3 volts above 50 volts, and ±4 volts from 20 to 50 volts.

Relay and Solenoid Response to Residual Voltage

Transfer in test A was made at 480 volts only (as would occur if a potential-transformer fuse opened, de-energizing relay 6), the 2,400-volt bus 1 having been manually transferred before test A began. Fig. 3 shows that the magnetic starters dropped out at about 0.4 second, when the voltage was approximately 50 per cent of normal, since the bus voltage fell abruptly to zero. One-tenth second later, relay 8 dropped out, point 1, but the transfer could not be completed until relay 6 timed out, tripping breaker F at 2.01 seconds, point 2. Breaker G closed 0.11 second later, completing the transfer. Drop-out time of relay 6 was thus increased over its 1.6-second dropout at zero voltage by the persistence of the partial voltage for the first half second. Table II shows all motors reaccelerated, so none of relays TD had timed out before transfer.

In test B, the 2,400-volt bus 1 was transferred with the 480-volt bus 1, but the simultaneous-tripping circuit tripped breaker F 0.13 second after breaker A opened. For this first 8th second, however, the two busses were connected through their transformers and a short length of 13.2-kv cable. The 2,300-volt motors, acting as generators for this short time, slowed voltage decay at 480 volts, as shown in Fig. 3. Except for this, the voltage decay was similar to test A, with relay 8 dropping out 0.13 second after the voltage dropped to zero. Since breaker F was already open, breaker G could close, completing the transfer at 0.75 second. Again, all 440-volt motors reaccelerated, but note that the system disturbance after the transfer lasted only 1/2 second in test B compared to 1¼ seconds in test A. The motors had not slowed as much in test B.

In test C as in test B, 2,300-volt motors were transferred with 440-volt motors, but circuit breakers C and F remained closed until tripped on undervoltage. Fig. 3 shows that the voltage on the 480-volt bus 1 was sustained by the 2,300-volt motors through the transformer connection, even though all starters had probably dropped out by 1 second, until relay 1 tripped breaker C at 2.2 seconds, which broke the connection. The voltage fell immediately to zero at this point. Relay 8 had dropped out at 1.85 seconds, but the slow decay of the voltage delayed timeout of relay 6 and tripping of breaker F until 2.55 seconds. By the time breaker G closed 0.1 second later, six of the relays TD had dropped out and the associated motors failed to reaccelerate; see Table II. Four other TD relays remained closed, presumably because of the manufacturing tolerances shown in Table I. The curves of test D showing a nearly identical transfer viewed from the 2,400-volt bus 1 correlate closely with test C, in particular confirming within 0.05 second the time at which breaker C opened. Two TD relays which dropped out in test C remained closed in test D.

Test C illustrates the difficulty encountered in co-ordinating unlike relays. The plungers of relays TD have a different equilibrium position for each voltage; above drop-out voltage, equilibrium is in the range of plunger overtravel. For gradual voltage drops, the plunger would lag behind equilibrium because of orifice restraints, the amount of lag depending on the rate of the voltage drop. If the drop-out voltage were reached in less than 5 seconds, for the particular relays used, some plunger overtravel would remain. Therefore operating time for relays TD is dependent on the rate of the voltage drop.

Table III. Data on Tests Opening Circuit Breaker B

Test Number	F	G	H	J
Oscillograph location				
2,400-volt bus 2	X	X	X	
480-volt bus 2				X
2,300-volt motors				
600 horsepower, 900 rpm*	N	N	R	R
450 horsepower, 3,600 rpm*	O	R	N	R
100 horsepower, 3,600 rpm	R	R	N	R
440-volt motors				
60 horsepower, 1,800 rpm	480-volt bus 2 fed through breaker G from source 1	R	480-volt bus 2 fed through breaker G from source 1	R
60 horsepower, 1,800 rpm		R		R
60 horsepower, 1,800 rpm		O		R
50 horsepower, 3,600 rpm*		R		R
50 horsepower, 3,600 rpm		R		R
40 horsepower, 3,600 rpm		R		R
25 horsepower, 3,600 rpm		R		R
15 horsepower, 3,600 rpm		R		R
15 horsepower, 3,600 rpm		O		R
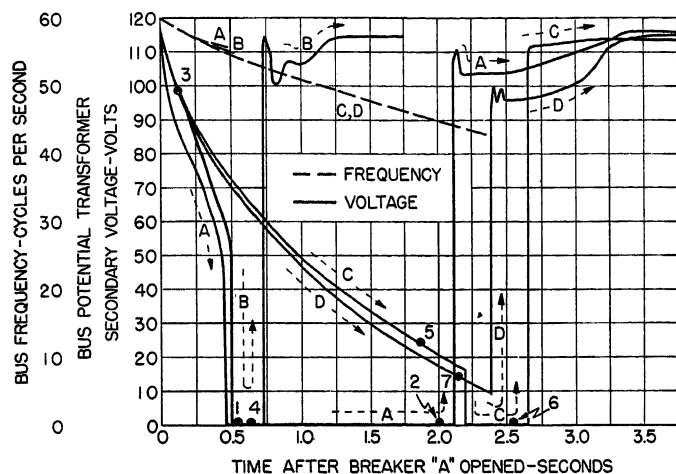
Fig. 3. Residual voltage and frequency and reaccelerating voltage during tests A, B, C, and D. Points 1 through 7 show time when residual-voltage relay dropped out and incoming circuit breakers opened

1—Relay 8 dropped out in test A
2—Breaker F opened in test A
3—Breaker F opened in test B
4—Relay 8 dropped out in test B
5—Relay 8 dropped out in test C
6—Breaker F opened in test C
7—Breaker C opened in test D
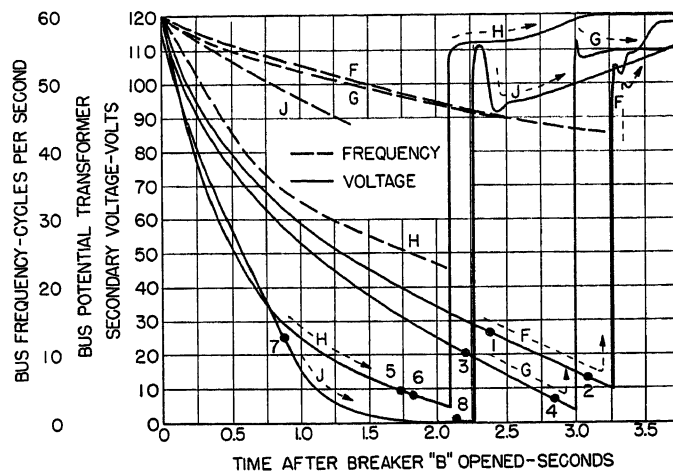
Fig. 5. Residual voltage and frequency and reaccelerating voltage during tests F, G, H, and J. Points 1 through 8 show time when residual-voltage relays dropped out and incoming circuit breakers opened

1—Breaker E opened in test F
2—Relay 5 dropped out in test F
3—Breaker E opened in test G
4—Relay 5 dropped out in test G
5—Relay 5 dropped out in test H
6—Breaker E opened in test H
7—Relay 9 dropped out in test J
8—Breaker H opened in test J

If the voltage were reduced abruptly from normal to zero, Table I shows that relays TD would take longer to operate then relay 6, yielding co-ordination. For slowly dropping voltage, the disk of relay 6 would not move until 93 volts were reached. If the voltage continued to drop uniformly, taking 5 seconds or more to reach the 54 to 66-volt dropout of relays TD (referred to 120 volts), relay 6 would time out before this voltage was reached, again yielding co-ordination. For intermediate rates of voltage drop, co-ordination is uncertain. Assuming that operating times of relays TD could be calculated for various shape voltage-time curves, several cases would have to be checked to be sure that relay 6 would operate first under all conditions. It ap-

pears that a safe expedient would be to choose a TD relay with drop-out time, for an abrupt drop from full voltage to zero, at least equal to the operating time of relay 6 at 50 per cent of normal bus voltage. (Relay 6 operates in 2.7 seconds at 60 volts.)

Residual voltage relay 8 played no part in tests A or C. In test B, however, it delayed transfer to 0.75 second at which time the various motor residual voltages are estimated to have been 35 to 40 per cent. Without relay, 8, the transfer would have completed in 0.25 second when the residual voltage was 70 per cent and starters were still closed.

Transfer in test E was made at 2,400 volts only, and with only two 2,300-volt motors running. As shown in Fig. 4,

relay 1 timed out and tripped circuit breaker C at 2.34 seconds. The residual voltage held relay 4 closed, however, until 3.23 seconds so that breaker D did not close until 3.44 seconds. Meanwhile, relays 3 had tripped both motor feeder breakers, the second at 3.35 seconds, so that neither motor reaccelerated. Assuming that the two motor breakers tripped at about the same time, it is seen that relays 3 co-ordinated with relay 1 as intended, timing out a full second later than relay 1. Their operation under such slowly decaying voltage, however, was slightly too rapid to permit successful transfer.

A conservative rule to prevent such failures would be to set the time of motor undervoltage relays at zero voltage about equal to the longest anticipated time for residual voltage to reach the drop-out point of the residual voltage relays. Assuming that these relays drop out at the desired 25 per cent of normal voltage, the tests indicate that time to reach this voltage can be as long as $2^{1}/_{4}$ seconds. In similar tests on power plant auxiliary motors, Johnson and Thompson[2] observed 2/3 second to reach 25-per-cent voltage, Gay[3] noted times as long as $2^{1}/_{2}$ seconds, Tevlin and Romzick (in an unpublished 1952 AIEE paper) report times over 1 second, and Backer, Barth, Huse, and Taylor[4] indicate times over 3 seconds. The times cited were obtained under extreme conditions of light load.

Fig. 4. Residual voltage and frequency during test E. Points 1 and 2 show time when

1—Incoming circuit breaker C opened
2—Residual relay 4 dropped out

However, nothing is lost by setting motor undervoltage relays for long enough times to hold in under such conditions. Times to reach 25-per-cent voltage can well exceed those noted in the foregoing if synchronous motors, capacitors, or very large induction motors are connected to the system. The author knows of two 3,000-horsepower induction motors, a 2-pole machine with a $4^1/_2$-second time constant, and a 4-pole machine with a $2^1/_4$-second time constant. It is suggested that time at zero voltage for motor undervoltage relays be set at 4 to 5 seconds.

It is seen that in test E relay 4 did not drop out until $1^1/_4$ seconds after the voltage passed its tested dropout of 26 volts. The other tests show similar results. This is attributed mainly to the drop in frequency along with voltage. As with most a-c instantaneous relays, the residual-voltage relays were highly inductive. Neglecting coil resistance, magnetic pull of such relays at a given armature position varies approximately as V^2/f^2, where V is voltage and f is frequency. Thus the voltage required to produce that constant pull just sufficient to hold the relay closed is approximately proportional to frequency, i.e., drop-out voltage is proportional to frequency. In addition, tests A and B show delays of about 0.1 second for these relays to close their normally closed contacts after being de-energized. Partial excitation remaining after the residual voltage passes the drop-out point would presumably delay drop-out further. The use of special residual voltage relays, with the drop-out voltage independent of frequency, will not only simplify determination of necessary time for motor undervoltage relays but will prevent unnecessarily long transfers with consequent prolonged system disturbance after transfer. (It is assumed that timing of the various induction-disk relays was increased because of frequency drop, since the torque for a given voltage is greater at lower frequency.)

Caution should be used in selecting residual voltage relays. Experience has shown a tendency of plunger-type a-c

relays in this service (as distinguished from clapper-type) to "hang up" after continuous energization at normal bus voltage, i.e., they have remained "picked up" when the voltage was reduced even to zero.

It should be noted that the discussions of relay co-ordination were based on all relays responding to the same voltage. If single-phase relays are used, the usual case, they should all be connected across the same phase as far as practicable, since a fault elsewhere on the system can cause asymmetrical voltages at the substation in question until the fault is cleared or the incoming circuit breaker at the substation trips on undervoltage.

Fig. 5 illustrates the effect of motor-generator action on the decay rate of the residual voltage and consequent relay response. Test F was similar to test E, transferring only two 2,300-volt motors. Notice that the voltage did not drop to zero in test F because one motor succeeded in reaccelerating; the other motor breaker was heard to trip just as the transfer was completed. As in test E, motor tripping resulted from relay 3 timing out before relay 5 dropped out, because of the slow decay of the residual voltage. With the same two motors running at 2,400 volts, the load of the 440-volt motors in test G (until their starters dropped out) speeded decay enough so that relay 5 dropped out before relays 3 operated, and both 2,300-volt motors reaccelerated. Some 440-volt motors failed to reaccelerate, however, indicating their TD relays dropped out as in tests C and D. Test H shows the rate of decay of a single 600-horsepower motor. Note the rapid drop in speed (frequency) causing rapid voltage decay. When this motor was transferred, in test J, with the same motors as in test G, voltage and frequency collapsed so fast that all motors reaccelerated.

To assure reacceleration of all motors in the substation described in the foregoing, the 2-second bottles of the TD mercury relays were replaced with 4-second bottles and 2,300-volt motor undervoltage-relay times were increased to approximately

$4^1/_2$ seconds at zero voltage. In addition, the residual voltage relays 4, 5, 8, and 9 were replaced with relays having a constant dropout of 30 volts from 25 through 60 cycles. After these changes, additional test transfers were made and all motors reaccelerated successfully.

Conclusion

In automatic transfers where motors reaccelerate, a simple means of limiting subtransient reaccelerating current is to delay transfer until the residual voltage drops to 25 per cent, and the following recommendations are made for applying relays:

1. Motor undervoltage protection must be time-delay type.

2. Motor undervoltage relays must have lower drop-out voltage and longer time at zero volts than undervoltage relays initiating transfer. If other than induction-disk type time-delay relays are used for motor undervoltage, it is suggested that their operating time at zero voltage be made equal to or greater than the operating time of the induction-disk transfer-initiating relay at 50 per cent of normal operating (not pickup) voltage.

3. The operating time of motor undervoltage relays at zero voltage should be at least 4 to 5 seconds, whether induction-disk type or not.

4. Relays employed to determine when residual voltage reaches 25 per cent should have a drop-out voltage independent of frequency from about 25 to 60 cycles.

5. To assure co-ordination, the transfer-initiating undervoltage relay, motor undervoltage relays, and residual voltage relay should all be connected across the same phase.

References

1. Transfer of Steam-Electric Generating Station Auxiliary Busses, D. G. Lewis, W. D. Marsh. *AIEE Transactions*, vol. 74, pt. III, June 1955, pp. 322–34.

2. Bus Transfer Tests on 2,300-Volt Station Auxiliary System, A. A. Johnson, H. A. Thompson. *Ibid.*, vol. 69, pt. I, 1950, pp. 386–90.

3. *Discussion* of reference 2 by F. W. Gay. Pp. 392–93.

4. Transfer Tests on Station Auxiliary Busses, L. E. Backer, Paul Barth, R. A. Huse, D. W. Taylor. *AIEE Transactions*, vol. 74. pt. III, 1955 (*Paper 55–92*).

Discussion

F. P. Brightman (General Electric Company, Schenectady, N.Y.):
Mr. Kelly's paper is an interesting and valuable addition to our store of knowledge as to the manner in which both the frequency and the voltage of motors decay in the interval immediately following disconnection from the power source.

The problem of how to prevent damage

to the machine, or alternatively serious reduction of the supply voltage caused by overload when motors are re-energized following a power outage comes up repeatedly:

1. When motors are transferred from one power source to another to maintain continuity of process operation.

2. When the power company has high-speed (15- to 20-cycle) reclosing on their supply lines to the industrial plant.

The problem is a very real one because as much as 20 times normal shaft torques have been measured during tests to determine the effect of re-energizing motors with high residual voltage out of phase with the power supply. Also, there have been reports from the field of cracked foundations and broken shafts resulting from the same cause. The term out-of-phase means sufficiently out of phase to make the resultant voltage, Fig. 1, appreciably greater than the source voltage. From a mechan-

ical-damage standpoint, induction motors over 200 or 300 horsepower and all synchronous motors should be protected against out-of-phase re-energization with high residual voltage. Incidentally, it should be kept in mind that the machine may not show any signs of distress due to the excessive torques created by out-of-phase re-energization for several repetitions of the operation, even though it is being damaged and will eventually show it. From the standpoint of system voltage disturbance caused by the excessive current drawn, all sizes and voltages of motors should be prevented from being re-energized out of phase with high residual voltage.

It is theoretically possible to disconnect and then re-energize a motor before it has gotten far enough out of phase with the power supply to cause trouble because of the excessive current drawn resulting from the high resultant voltage created by the combination of the supply voltage and the motor residual voltage. Practically, it is exceedingly difficult if not impossible to do so in many cases, because the permissible outage time is less than the minimum possible reclosing or transfer time with existing breakers and relays. Granting this is true, it becomes necessary to prevent re-energization of the motors with sufficient out-of-phase residual voltage to make trouble. The scheme described in Mr. Kelly's paper appears to be a very satisfactory solution of the problem as far as automatically transferred motors are concerned. It accomplishes the objective of keeping the machine and system out of trouble due to too much residual voltage on the motors when they are re-energized, and also insures that the motor-driven machinery will be restored to service in the minimum safe time.

I would like to re-emphasize Mr. Kelly's statement that undervoltage relays used in such a scheme should be designed with the necessary characteristics to make them operate independently of frequency. Otherwise, they will not operate at the expected values of voltage.

It is interesting to note from Mr. Kelly's charts that the motors and their driven pumps, etc., did not slow down as much or as rapidly as might have been expected. This may be because the motors were driving pumps and the fluid flow in the system had sufficient inertia to help hold up the unit's speed by driving the pump as an hydraulic turbine.

Motors that are on circuits that will be re-energized because of high-speed reclosing of power company breakers should be protected by being disconnected from the line prior to its re-energization by high-speed (15 to 20 cycles) reclosing of circuit breakers. At present, the best means of obtaining the signal to disconnect the machine is from a high-speed underfrequency relay. However, even these may not be fast enough in some cases. If this appers likely, a dead-line-check (undervoltage) relay can be used at the power company end of the line to prevent automatic reclosing of the power company breaker until the motor residual voltage has dropped to a safe value. This, of course, requires co-operation between the power company and the user.

Mr. Kelly is to be commended for making this information available to the industry. More information on both phases of the subject is needed, and it is hoped that other people will follow his example.

The following items of information, which I obtained from representatives of a power company in the Southwest may be of interest.

In one case, an oil refinery customer with a 4,000- to 5,000-kw load (about 30-per-cent synchronous motors) made some tests in conjunction with the power company to determine whether or not the plant could be disconnected and re-energized without shutting down the machines due to under-voltage caused by the excessive current drawn when the system was re-energized. The tests indicate that they could re-energize the system successfully and with no detected mechanical injuries to machines.

The same power company people said that immediate high-speed reclosing of power lines supplying large pipe-line motors was standard practice and no difficulties had been reported. In this particular case, it is possible that the inertia of the moving fluid in the pipe was sufficient to help maintain the speed of the motors by driving the pumps. If this surmise is true, it might be that the "motoring" effect was sufficient to keep the motors from getting far enough out of phase to be troublesome regardless of how high the residual voltage was.

D. W. Taylor (Public Service Electric and Gas Company, Newark, N. J.): It is interesting that more and more people are studying the problems connected with the transfer of load from one source to another. With more test data becoming available and more thought being given to the subject, better solutions will result.

Our experience has been entirely with the station auxiliaries used with large turbine generators. While the residual-voltage method of inaugurating the transfer was satisfactory in many cases, we were confronted with the necessity of reducing the time of transfer to 3/4th second or less in our latest installation and resorted to controlled angle reclosing at higher residual voltages. This method is in satisfactory operation and can be applied where consistent bus motor loading is practical. However, Mr. Kelly points out that industrial motor applications vary to the extent of exclusion of this practice for bus transfers. If necessary, although at additional cost, individual motors could be transferred by the controlled angle method.

We have had the same difficulty with residual voltage relays as described by Mr. Kelly. Plunger-type a-c relays have failed to drop out even at zero voltage. They have been replaced by the clapper-type relays and these, in general, vary considerably on repeat operations. Relays set for 25-per-cent dropout resulted in operation at 10-, 15-, or 18-per-cent residual voltage. We have, therefore, set these relays at a higher value, namely, around 35 per cent to accomplish a 25-per-cent reclosing. This, I believe, is because of an a-c dropout relay's operation is inaccurate under about a 50-per-cent setting. We are now testing a d-c relay using rectifiers to

convert the alternating current to direct current and hope to get more consistent results.

In some installations it is not essential to have accuracy in the drop-out point of the residual voltage relay, as a zero dropout may be just as good as the 25-per-cent dropout. However, this increases the time of the throwover, and in several of our installations the time was critical and reasonable accuracy was desirable.

We believe that the demands for continuous power flow in connection with thermochemical and mechanical processes will increase and that the problem presented in this paper will be of general interest to a great many industries. We congratulate Mr. Kelly on his pursuit of this subject and his presentation of the problem.

A. R. Kelly: Opinions differ on the risk incurred by transferring with no regard for residual voltage. Mr. Brightman reports broken shafts and foundations, and the discussion of reference 1 lists a winding failure. Yet, both Mr. Brightman and other discussers of reference 1 list installations which successfully re-energize with neither intentional delay for residual voltage decay nor any provision for synchronizing. In any case, the risk of motor tripout is quite real. We first began using delayed transfer after one in which several motors tripped off on reaccelerating inrush current. Just recently a miscalibrated motor overcurrent relay was discovered after a transfer. The motor had previously been started a number of times with an instantaneous overcurrent setting of 140-per-cent locked rotor current; it tripped out on transfer, however, and the setting was then raised to the specified 180 per cent.

Mr. Brightman points out that the same problems exist for industrial systems supplied from utility lines with reclosers. As Mr. Brightman implies, even starters with "instantaneous" undervoltage protection may be held in through a reclosure by the residual voltage of their motors, and artificial means are required to disconnect them before reclosure. Mr. Brightman mentions high-speed underfrequency tripping. Since the slower induction-disk underfrequency relay is surely too slow for this purpose, the high-speed balance-arm type would be necessary. However, I believe a calibrated variable-frequency source is required for setting these high-speed relays, and, as Mr. Brightman points out, they may still be too slow. It appears that the utilities should take the responsibility for warning those customers with large motors if the system uses reclosers, unless they delay reclosure until voltage decays.

Both Mr. Brightman and Mr. Taylor discussed the necessity of selecting residual voltage relays designed for the purpose. I can corroborate Mr. Taylor's experience of drop-out voltage variation in a-c relays. Dropouts may vary by considerably more than can be explained by frequency change. We now use either of two relays, both d-c, fed from 3-phase rectifiers, which are offered as meeting the following specification:

"Residual voltage relay, dropout adjust-

able from 30 to 45 volts, suitable for continuous operation at 120 volts without tendency to stick in the energized position. For any dropout voltage setting: (1) dropout shall not vary more than 2 volts for frequencies between 25 and 60 cycles; (2) pickup shall not exceed 95 volts; (3) hot dropout (after continuous operation at 120 volts) shall be within 4 volts of cold dropout (after prolonged de-energization

and then momentary application of 120 volts); and (4) dropout voltage shall be independent of prior applied voltage or else the relay shall be prominently tagged 'Raise voltage to 115V before calibrating dropout.' The relay shall be complete within a semi-flush mounting case. Electrolytic capacitors shall not be used. Normally-closed contact circuit shall remain closed with relay withdrawn."

These relays can be set easily and are simple in principle and construction. This is of prime importance for industrial relays. Our newly standardized automatic transfer circuit, developed after experience with the many unsuspected misoperations possible with such circuits, is still relatively simple and uses no device more complicated than an induction-disk over-current relay.

Extension of Continuous-Data System Design Techniques to Sampled-Data Control Systems

G. W. JOHNSON
ASSOCIATE MEMBER AIEE

D. P. LINDORFF
ASSOCIATE MEMBER AIEE

C. G. A. NORDLING
NONMEMBER AIEE

THE purpose of this paper is to extend the utility of certain synthesis techniques, which have proved useful in the design of continuous feedback systems, to the study of sampled-data systems. In recent publications,[1,3,5] the z-transformation has been utilized to advantage in simplifying the analysis of sampled-data systems. However, several useful techniques applicable to the design of continuous-data systems have not been made available for the design of sampled-data systems. In this paper, by means of an extension of the concepts embodied in the z-transformation, the application to sampled-data systems of root-locus plots, asymptotic frequency plots, and Routh's criterion is shown to be feasible. Through the use of these techniques, a digital program giving the desired system compensation is readily synthesized.

The use of the z-transformation permits the system response to be specified only at the sampling instants. Hence, to permit the application of the aforementioned design techniques, a linearized system response is defined which agrees with the actual system response at the sampling instants. In any specific problem it is

desirable to know how accurately the linearized system response defines the actual response. For this purpose, a simple relationship is derived for determining the actual system response in any time region of interest. In the following the foregoing concepts are discussed and, for the purpose of clarification, these concepts are applied to a specific problem. A possible digital program is suggested as a means of compensation.

Nomenclature

INDEPENDENT VARIABLES

t = real variable
z = transform variable = ϵ^{sT}
w = transform variable = $\dfrac{z-1}{z+1}$ = $u+jv$

FUNCTIONAL NOTATION

$f(t)$ = time function
$F(s) = \mathcal{L}f(t)$
$\delta_T(t)$ = unit impulse train
$F^*(t) = f(t)[\delta_T(t)]$
$F^*(s) = \mathcal{L}f^*(t)$
$F^*(z) = F^*(s)$ for $z = \epsilon^{sT}$
$F^*(w) = F^*(z)$ for $z = \dfrac{1+w}{1-w}$

SYSTEM PARAMETERS

K_v = velocity constant
τ = time constant
$K_u = K_v\tau$ = dimensionless velocity constant
T = sampling period
$\omega_0 = \dfrac{2\pi}{T}$ = sampling frequency
$\alpha = \dfrac{T}{\tau}$ = dimensionless sampling period

TERMS USED TO CHARACTERIZE SYSTEM STABILITY

σ = decrement factor
ω = damped natural frequency

$\zeta = -\cos\left(\tan^{-1}\dfrac{\omega}{\sigma}\right)$ = damping ratio
$\omega_n = |s|$ = undamped natural frequency
$M_p = \left|\dfrac{C}{R}\right|_{max}$ = maximum ratio of the closed-loop frequency response
ω_m = the frequency at which $\left|\dfrac{C}{R}\right|_{max}$ occurs

Inverse z-Transformation to Obtain a Continuous Response

The primary purpose of this section is to develop a process of real inversion for the z-transformation that will yield a continuous response agreeing with the actual system response at the sampling instants. It will be shown that the design of sampled-data systems can be accomplished in terms of the location of the poles of the response function in the z-plane in a manner analogous to that used in the design of continuous systems. Although the continuous response so obtained is only exact at the sampling instants, it will be demonstrated later that it may be used to characterize the actual system response by a relatively simple analytical technique. The basic concepts of the z-transform method of analyzing sampled-data systems, as presented by Ragazzini and Zadeh,[1] will be repeated here to clarify the concept of a real inversion to yield a continuous response. The z-transform technique will be developed for the system shown in Fig. 1, which is an open-ended linear system receiving pulsed data. The notation used is in agreement with recent publications[1] and is listed in the Nomenclature.

The sampling process is represented by a switch which closes momentarily every T seconds. The time during which the switch is closed is assumed to be short compared to the sampling period. Thus, the output of the sampler at any sampling instant is a finite pulse whose magnitude is equal to the input at the sampling instant $t = nT$. If the time during which the switch is closed is short compared to the time constants of the system receiving the sampled data, the finite pulses may be replaced by impulses whose magnitudes equal the amplitudes of the sampled
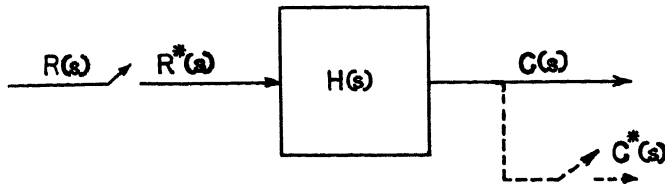
Fig. 1 (above).
Pulsed linear system

Fig 2 (right). Symmetrical distribution of strips in the s-plane due to periodicity of $C^*(s)$

data at the sampling instants. Thus, the output of the sampler can be simulated as a unit-impulse train, amplitude-modulated by the magnitude of the sampled signal. The unit-impulse train may be represented as

$$\delta_T(t) = \sum_{n=0}^{n=\infty} \delta(t-nT) \qquad (1)$$

which allows the output of the sampler to be written as

$$r^*(t) = r(t)[\delta_T(t)] = \sum_{n=0}^{n=\infty} r(nT)\delta(t-nT) \qquad (2)$$

The Laplace transform of $r^*(t)$ may be obtained by transforming the infinite summation in equation 2, term by term, by the use of real translation.[2] This process gives

$$R^*(s) = \sum_{n=0}^{n=\infty} r(nT)\epsilon^{-nTs} \qquad (3)$$

Equation 3 illustrates the well-known fact that $R^*(s)$ may be expressed as a power series in ϵ^{-Ts} where the coefficient of ϵ^{-nTs} is the value of $r(t)$ at $t=nT$.

$R^*(s)$ may be obtained in an alternate form. Since $\delta_T(t)$ is a periodic function it may be expanded in a Fourier series in exponential form as

$$\delta_T(t) = \frac{1}{T} \sum_{n=-\infty}^{n=\infty} \epsilon^{jn\omega_0 t} \qquad (4)$$

where $\omega_0 = 2\pi/T$ is the sampling frequency. Therefore, $r^*(t)$ may be written as

$$r^*(t) = \frac{r(t)}{T} \sum_{n=-\infty}^{n=\infty} \epsilon^{jn\omega_0 t} \qquad (5)$$

The infinite summation in equation 5 may be transformed, term by term, by the use of complex translation[2] to yield

$$R^*(s) = \frac{1}{T} \sum_{n=-\infty}^{n=\infty} R(s+jn\omega_0) \qquad (6)$$

From equation 6 it is apparent that $R^*(s)$ is periodic in s with period $j\omega_0$. This periodicity may be stated mathematically as

$$R^*(s) = R^*(s+jm\omega_0) \qquad (7)$$

where m is any integer, positive or negative. It should be emphasized at this

time that $r^*(t)$ is not a true simulation of the output of the sampler and that the transfer function of the components following the sampling process should be assigned on the basis that they are receiving impulses containing only the amplitude of the sampled signal.

With reference to Fig. 1, the Laplace transform of the response may be written as

$$C(s) = R^*(s)H(s) \qquad (8)$$

If the response $c(t)$ is sampled in unison with $r(t)$, then an expression for $C^*(s)$ is obtained which is a rational fraction in ϵ^{Ts}. This property will be demonstrated when the z-transformation is introduced. Since $C^*(s)$ may be obtained as a rational fraction, the advantage of considering the response to be sampled is apparent.

From equations 6 and 8 it is seen that $C^*(s)$ may be written as

$$C^*(s) = \frac{1}{T} \sum_{n=-\infty}^{n=\infty} R^*(s+jn\omega_0)H(s+jn\omega_0) \qquad (9)$$

Since $R^*(s) = R^*(s+jn\omega_0)$, equation 9 becomes

$$C^*(s) = \frac{R^*(s)}{T} \sum_{n=-\infty}^{n=\infty} H(s+jn\omega_0) \qquad (10)$$

which by reference to equation 6 may be written as

$$C^*(s) = R^*(s)[H^*(s)] \qquad (11)$$

Equation 11 represents one of the basic theorems of the algebra of pulsed linear systems and will be referred to in the following section where its application to pulsed systems employing feedback is dis-

cussed. Since both $R^*(s)$ and $H^*(s)$ are periodic functions of s with period $j\omega_0$, $C^*(s)$ also has a period $j\omega_0$. Thus the poles and zeros of $C^*(s)$ are periodic in distribution throughout the s-plane. Fig. 2 shows the s-plane divided into strips parallel to the real axis and ω_0 in width. The first strip was chosen symmetrical with respect to the real axis to obtain a symmetrical distribution of poles and zeros in that strip. It must be kept in mind that $C^*(s)$ defines the response only at the sampling instants and, therefore, the real inversion may be considered to be any time function which yields the correct response at the sampling instants.
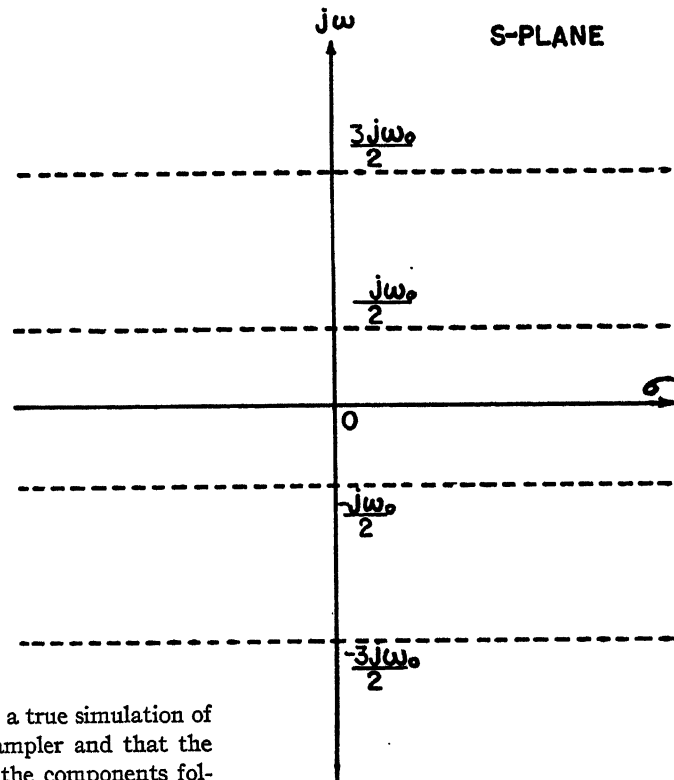
At this point the z-transformation will be introduced as a change of the independent variable given by $z=\epsilon^{sT}$. As is well known,[1] the utility of this transformation is that the sampled response may be placed in closed form as a ratio of polynomials in z. Therefore $C^*(z)$ contains only a finite number of poles in z, while $C^*(s)$ has a nonfinite number of poles because of its periodicity.

With reference to Fig. 1, $C^*(s)$ is equivalent to $C^*[(1/T)\ln z]$ which will be referred to as $C^*(z)$ for notational convenience. This equivalence may be stated in equation form as

$$C^*(z) = C^*(s) \qquad (12)$$

where $z=\epsilon^{sT}$. When the solution is effected in the z-domain, it is convenient to use the real inversion integral[3] expressed as a contour integral in the z-plane. This

Fig 3. Basic feedback system

integral is derived in Appendix I and the result is

$$c(mT) = \frac{1}{2\pi j} \oint_\Gamma C^*(z) z^{m-1} dz \qquad (13)$$

where the contour Γ encloses all the singularities of the integrand.

It should be noted that in equation 13 the response is not readily specified in terms of a damping ratio ζ and natural frequency ω_n. The advantages in using ζ and ω_n as time-domain specifications are well known for linear continuous systems and have served as simple relationships between the time and complex-frequency domains. Equation 13 may be written as

$$c(mT) = \frac{1}{2\pi j} \oint_\Gamma C^*(z)_\epsilon \left(\frac{1}{T} \ln z\right)(m-1)T dz \qquad (14)$$

If mT is replaced by t in the solution of equation 14, a continuous response is obtained which will agree with the sampled response at the sampling instants. The continuous response so obtained will be called the linearized system response $c_1(t)$, where $c_1(t) = c(mT)$ for $t = mT$. If $-\pi < \arg z \le \pi$, the linearized system response will be defined by the poles of $C^*(s)$ in the first strip in the s-plane, since $s = (1/T) \ln z$. Thus the linearized system response may be characterized by the damping ratio and natural frequency of the dominant poles of $C^*(s)$ in the first strip in the s-plane. The use of this response to characterize the actual system response at other than the sampling instants will be discussed in detail in a subsequent section.

Application of z-Transforms to Closed-Loop Systems

The material in the preceding section has set forth the mathematical concepts which serve to characterize the sampled-data signal. Up to this point, no mention has been made of closed-loop systems in

which the sampling process takes place within the loop. Since the remainder of this paper will be concerned with the development of techniques which are useful in studying the stability of closed-loop sampled-data systems, it will be helpful to define a basic feedback system as shown in Fig. 3. Because the z-transformation implies the physical process of sampling, the desired quantity $C^*(z)$ has been shown to be derived from the continuous output by a sampling process.

The equations defining the block diagram in Fig. 3 are given by

$$C(s) = G(s)E^*(s) \qquad (15)$$

$$R(s) - C(s) = E(s) \qquad (16)$$

To obtain $C^*(s)$ and, hence, $C^*(z)$, use is made of the theorem stated by equation 11. This gives

$$C^*(s) = [G(s)E^*(s)]^* = G^*(s)E^*(s) \qquad (17)$$

Similarly

$$R^*(s) - C^*(s) = E^*(s) \qquad (18)$$

The elimination of $E^*(s)$ from equations 17 and 18 yields

$$C^*(z) = R^*(z) \frac{G^*(z)}{1 + G^*(z)} \qquad (19)$$

It should be noted that although the foregoing expression for $C^*(z)$ is readily obtained, in more complex systems it is necessary to resort to special techniques to solve for the desired quantity. In Appendix II, the algebra is shown pertaining to a sufficiently complicated system to illustrate the general method of attack.

Root Loci in the z-Plane

Application of root-locus techniques to sampled-data systems serves a unique purpose in the sense that, in addition to focusing attention on the effect of gain variations upon the system dynamics, it further serves to emphasize the effect of variations in sampling period upon the dynamic response. In this section the means by which root-locus techniques can be applied to sampled-data systems will be discussed. The procedures by which the locus of the roots can be derived will be seen to be identical in principle to the case of continuous systems and will not be considered herein. However, certain features of the root-locus plots peculiar to sampled-data systems are deserving of comment and these will be discussed with reference to an example in a subsequent section.

With reference to Fig. 3, the response ratio is given in terms of the z-transform variable as

Fig. 4. Conformal mapping according to the relationship $z = \epsilon^{sT}$

Johnson, Lindorff, Nordling—Continuous-Data System Design

Fig. 5. Conformal mapping of ζ, σ, and ω

$$\frac{C^*(z)}{R^*(z)} = \frac{G^*(z)}{1+G^*(z)} \qquad (20)$$

Since $G^*(z)$ is the quotient of two polynomials in z with an associated gain factor, it follows directly that, for a given sampling period, the zeros of $1+G^*(z)$ as a function of the gain factor can be determined by finding the values of z which make $G^*(z) = -1$. For the purposes of design procedure, it is convenient to consider only the poles of $C^*(s)$ within the first strip as characterizing the so-called linearized system response $c_1(t)$. Fig. 4 shows the conformal mapping of the first strip in the s-plane into the unit circle in the z-plane.

By transforming contours of constant ζ (damping ratio), constant σ (decrement factor), and constant ω (damped natural frequency) from the s-plane into the z-plane, the system performance as related to $c_1(t)$ can be determined directly by inspection of poles in the z-plane. The graph in Fig. 5 illustrates the significant features of such a plot, with the associated defining equations for ζ, σ, and ω.

Application of the root-locus method to the design of a more complex system than represented in Fig. 3 will be illustrated in a subsequent section.

Application of Conventional Techniques Through the Use of a Bilinear Transformation

With reference to Fig. 3, it was demonstrated that the closed-loop transfer ratio in terms of the values of the signals at the sampling instants may be written as

$$\frac{C^*(s)}{R^*(s)} = \frac{G^*(s)}{1+G^*(s)} \qquad (21)$$

Thus by plotting $G^*(s)$, for $s=j\omega$, a closed-loop frequency response may be obtained from which system stability may be determined. Two methods have been developed[1,4] from which $G^*(j\omega)$ may be plotted. Linvill[4] plots the $G^*(j\omega)$ locus by letting $s=j\omega$ in the infinite summation for $G^*(s)$. Thus the number of significant terms in the summation is dependent on the filter characteristics of any specific system. Ragazzini and Zadeh[1] obtain

the open-loop frequency plot directly from the closed-form expression of $G^*(z)$ by plotting the $G^*(z)$ locus for $|z|=1$, corresponding to $s=j\omega$. Neither of the foregoing methods can utilize the asymptotic plotting techniques which so greatly simplify the construction of the frequency-response characteristics.

Consider a change of independent variable given by $z=(1+w)/(1-w)$. The use of such a bilinear transformation preserves $G^*(w)$ as a ratio of polynomials. However, the unit circle in the z-plane corresponding to $s=j\omega$ transforms into the entire imaginary axis in the w-plane, as shown in Fig. 6. Thus the open-loop frequency locus may be constructed for $G^*(w)$ by letting $w=jv$. The equivalence of the loci obtained by all three methods is shown in equation form where by definition

$$G^*(z) = G^*(w) = G^*(s)$$

where

$$z = \frac{1+w}{1-w} = \epsilon^{sT} \qquad (22)$$

A few of the immediately apparent advantages to be gained by making the w-transformation are as follows:

1. Use of asymptotic plotting techniques to obtain a frequency response from which relative stability may be predicted by use of the Nichols chart.

2. Design of digital compensation in terms of lead and lag networks in the w-plane.

3. Application of Routh's criterion to
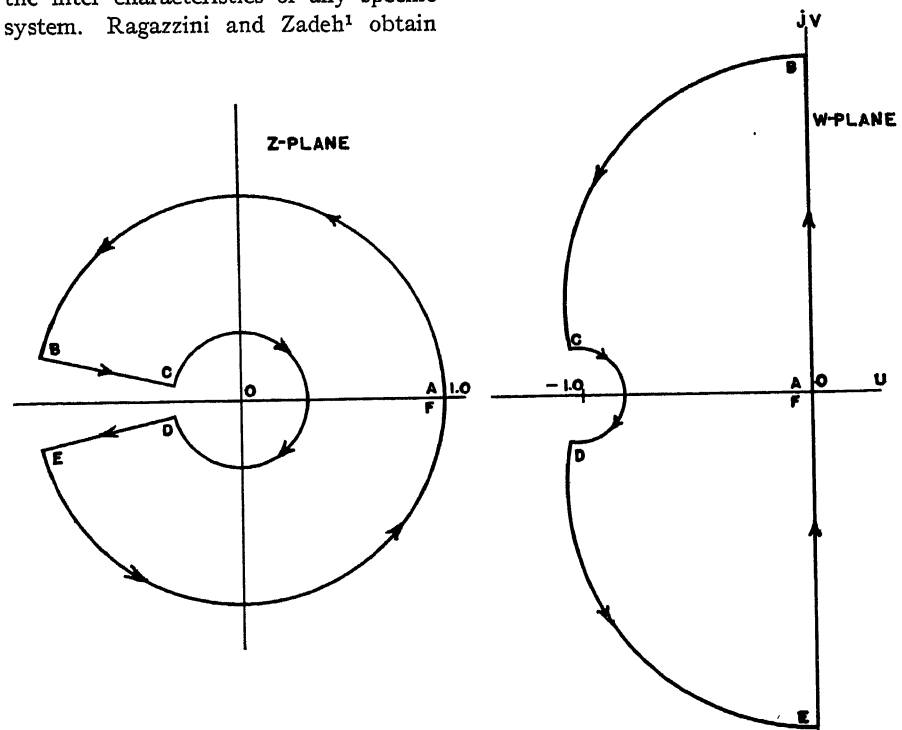
Fig. 6. Conformal mapping according to the relationship $w = z-1/z+1$

determine absolute stability of sampled-data control systems.

Each of the foregoing statements will be clarified in a later section by application of the w-transformation to a specific example. An abbreviated listing of w-transforms is shown in Table I.

Study of More Complex Systems

In the previous sections pertaining to root-locus and frequency-locus plots, reference was made to the basic system shown in Fig. 1. As in stability studies of continuous systems, both of the aforementioned techniques may be used to advantage in stability studies of more complex systems involving sampled data. To clarify this point, it is simply necessary to recognize that the loop stability is specified by the zeros of the characteristic equation, namely (1+loop transmission). As is customary in the study of continuous systems, stability can be specified in terms of a damping ratio ζ or in terms of the magnitude of the resonant peak in the closed-loop frequency response (M_p criterion), where it is tacitly assumed that the loop stability is characterized by a dominant pair of quadratic roots. By the use of the linearized system response, the foregoing concepts may be applied to more complex sampled-data systems.

The possibility exists of introducing a pole in the closed-loop transfer function which is not specified by the zeros of (1+loop transmission). This situation is a special case which can arise in more complex continuous- or sampled-data systems. As a simple illustration, consider the closed-loop transfer function given by

$$\frac{C}{R} = \frac{G}{1+GH} \qquad (23)$$

To apply the M_p criterion by use of the Nichols chart, it is necessary to rewrite equation 23 in the form

$$\frac{C}{R} = \frac{1}{H} \frac{GH}{1+GH} \qquad (24)$$

Normally, the poles of C/R will be given by the zeros of $1+GH$. However, if a zero of H is identical to a pole of G, a perfect cancellation in the numerator of equation 24 will occur and a corresponding pole will appear in C/R because of the factor of $1/H$. Therefore, although the M_p criterion can be used to specify the loop stability by referring to $GH/(1+GH)$, the M_p criterion can be relied upon to specify the transient response of the system only by referring to the exact expression for the frequency response

given by $1/H[GH/(1+GH)]$. Similarly, when using root-locus techniques in evaluating the system dynamics, the poles of C/R that result from the $1/H$ factor must also be considered when such perfect cancellation occurs.

Though experience may well justify the application of these concepts to a specific system study, it is wise to be aware of the possible sources of error. One source of error is caused by the approximation in assuming the complex system response to be characterized by a dominant quadratic mode. This assumption is by direct analogy related to the restraints imposed upon continuous systems which are well known in the field of automatic control and, as such, will not be considered further here. The second source of error is caused by the inexactness in representing the actual system response by a linearized system response. Since a method of evaluating this source of error is important in any system study, whether simple or complex, the following section will be concerned with a general method for evaluating the degree of accuracy obtained.

Exact Response Between Sampling Instants

A relatively simple analytical technique is outlined that will allow the evaluation of the exact system response during any sampling period. It is to be pointed out, however, that such a technique will not lend itself readily to synthesis procedures but will merely serve to indicate the validity of the use of the linearized system response to characterize the actual system response between sampling instants and thus allow root-locus and frequency-response techniques to be used in the synthesis of sampled-data control systems.

With reference to Fig. 3, it is seen that the Laplace transform of the actual sys-

tem response may be written as

$$C(s) = E^*(s)G(s) \qquad (25)$$

from which $c(t)$ may be obtained by the process of real convolution as

$$c(t) = \int_0^t \epsilon^*(\tau)g(t-\tau)d\tau \qquad (26)$$

where $g(t) = \mathcal{L}^{-1}G(s)$. Equation 26 may be written as

$$c(t) = \sum_{n=0}^{n=\infty} \left\{ [r(nT) - c(nT)] \left[\int_0^t \delta(\tau - nT)g(t-\tau)d\tau \right] \right\} \qquad (27)$$

which in turn may be expressed as

$$c(t) = \sum_{n=0}^{n=m} [r(nT) - c(nT)][g(t-nT)] \qquad (28)$$

for $mT \leqslant t \leqslant (m+1)T$.

With reference to equation 28, $r(nT)$ is immediately available from the form of the excitation function and $c(nT)$ may be determined by the use of z-transforms. Thus, $c(t)$ may be determined during any sampling period by effecting the finite summation represented in equation 28. It should be pointed out that in practice $c(t)$ only needs to be calculated for a relatively few sampling periods to determine how accurately the linearized system response characterizes the actual system response between sampling instants. For instance, if peak overshoot in step-function response is used as a criterion for stability, the actual system response need be calculated only during one sampling period as will be demonstrated in the following section with respect to a specific example.

Example

The techniques developed in the preceding sections will now be demonstrated with respect to a specific example. The block diagram shown in Fig. 7 represents the system to be used for illustrative pur-
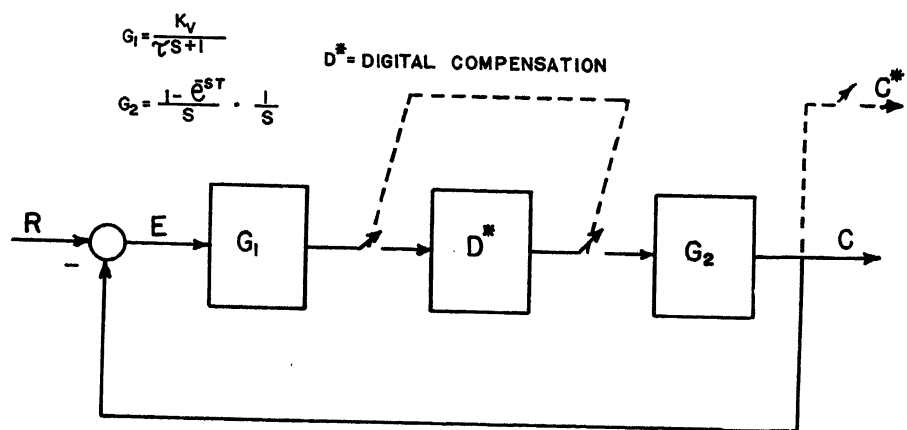
Fig. 7. System used as an illustrative example

poses. With reference to Fig. 7, $G_1(s)$ and $G_2(s)$ represent the conventional linear transfer functions of the continuous portions of the system where $G_2(s)$ includes a zero-order hold circuit. $D^*(z)$ represents a linear digital program that will be used as a means of compensation.

The block diagram algebra associated with this system is developed in Appendix II where it is shown that the z-transform of the response function may be written as

$$C^*(z) = \frac{RG_1^*(z)G_2^*(z)D^*(z)}{1+G_1G_2^*(z)D^*(z)} \qquad (29)$$

where G_1, G_2, and D are defined in Fig. 7.

For the case where $r(t) = U(t)$, $C^*(z)$ may be written in dimensionless form as

$$C^*(z) =$$

$$\frac{\alpha K_u\left(\dfrac{(1-\epsilon^{-\alpha})z}{(z-1)^2(z-\epsilon^{-\alpha})}\right)D^*(z)}{1+K_u\left(\dfrac{(\alpha-1+\epsilon^{-\alpha})z+(1-\epsilon^{-\alpha}-\alpha\epsilon^{-\alpha})}{(z-1)(z-\epsilon^{-\alpha})}\right)\times D^*(z)} \qquad (30)$$

where

$$\alpha = T/\tau$$
$$K_u = K_s\tau$$

The system shown in Fig. 7 will now be analyzed by the root-locus and frequency-response techniques so that a desirable system stability may be selected. A simple digital program will be synthesized to typify a method of compensation.

Root Locus Study of Uncompensated System

Having derived the expression for the response $C^*(z)$, the root-locus plot in the z-plane can be derived directly by plotting the zeros of the characteristic equation $1 + G_1G_2^*(z)$ since for the uncompensated system these zeros alone characterize the transient response. It is of interest to plot the root loci in the conventional manner where the system gain is the independent variable, and the dimensionless sampling period α is a fixed parameter.

As is well known, the zeros and poles of $G_1G_2^*(z)$ are the terminal points of the closed-loop roots. For the system under consideration, $G_1G_2^*(z)$ is given by

$$G_1G_2^*(z) = (\alpha-1+\epsilon^{-\alpha})K_u\times$$
$$\left(\frac{z+\dfrac{1-(\alpha+1)\epsilon^{-\alpha}}{\alpha-1+\epsilon^{-\alpha}}}{(z-1)(z-\epsilon^{-\alpha})}\right) \qquad (31)$$

It can be shown that in the foregoing expression, the zero of $G_1G_2^*(z)$ falls between 0 and -1 for $0<\alpha<\infty$. The root loci plotted in Fig. 8 for $\alpha=1$ are

seen to originate as two positive real roots becoming complex conjugates, and ultimately turning into a pair of negative real roots. It is of interest to note that the locus of the complex roots describes a circle with a center at the zero of $G_1G_2^*(z)$.

The root loci as plotted in Fig. 8 characterize the linearized system response as a function of K_u for a given α. For any given dimensionless sampling period, the gain can be determined which gives rise to a desired damping ratio for the oscillatory roots. In this example, for $\alpha=1$, a value of $K_u=0.63$ results in a damping ratio $\zeta=0.4$ as shown. The damped natural frequency of the roots so obtained is seen to be $\omega=(0.22)(\omega_0/2)$. Hence the optimum bandwidth of the system, namely $\omega_0/2$, has not been realized

Fig. 8 (right). Root-locus plot of uncompensated system for $\alpha = 1$
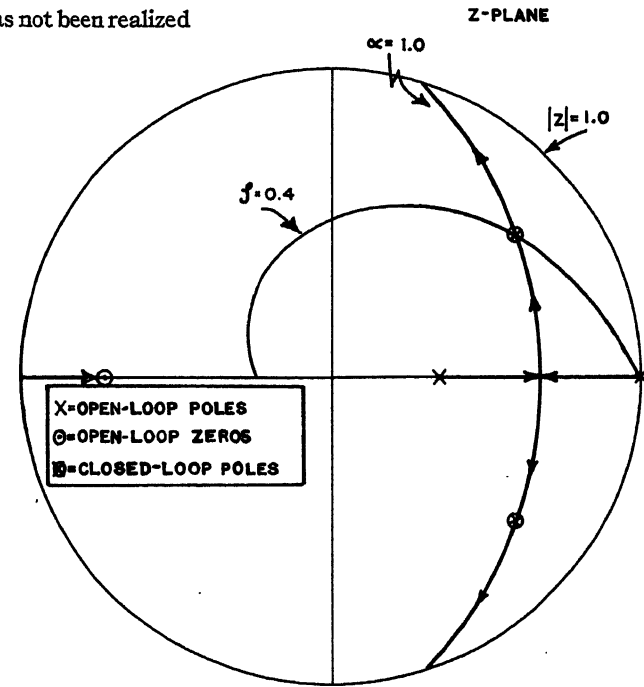
Fig. 9 (below). Attenuation and phase plot for uncompensated and compensated systems

and compensation is required if the speed of response is to be improved.

In the following, the foregoing results will be correlated with those obtained by application of conventional asymptotic-frequency plotting techniques, and a digital program will be synthesized as a possible means of compensation.

Application of Frequency Response Techniques

By the use of the transformation $z=(1+w)/(1-w)$, it has been shown that the real-frequency plot of $G^*(j\omega)=G^*(jv)$, where $\omega=(2/T)(\tan^{-1}v)$, can be obtained by the use of standard asymptotic plotting

Johnson, Lindorff, Nordling—Continuous-Data System Design

techniques. Using this transformation, and letting $\alpha = 1$, equation 31 yields

$$G_1 G_2{}^*(w) = \frac{K_u}{2} \frac{(1+w/6.1)(1-w)}{w(1+w/0.46)} \qquad (32)$$

The log-modulus frequency plot of equation 32 for $w = jv$ is shown in Fig. 9. Note that the foregoing transfer function is nonminimum phase in w, though this offers no serious obstacle to the plotting technique. Application of the 45-degree phase-margin criterion for stability yields a gain of $K_u = 0.63$, consistent with that obtained previously from the root-locus plot. This in turn is found to result in an $M_p = 2.5$ decibels, and a bandwidth $\omega_m = (0.186)(\omega_0/2)$.

It is now of interest to consider possible means of compensation by synthesizing a digital program $D^*(w)$, as shown in Fig. 7. Since the digital program is in itself a discrete filter, it is permissible to write the compensated loop-transfer function as $G_1 G_2{}^*(w) \, D^*(w)$. It should be noted that, if $D^*(w)$ has no poles in the right-half plane or on the jv axis, it will correspond to a stable computer program since $|z| < 1$ maps into the left-half w-plane.

For the purposes of illustration a lag network is employed as a means of compensation to realize an improvement in system gain. A conventional passive lag network in w results in

$$D^*(w) = \frac{1 + \dfrac{w}{0.046}}{1 + \dfrac{w}{0.0046}} \qquad (33)$$

The compensated asymptotic frequency locus is plotted in Fig. 9. For the same value of phase margin as used previously, the system gain has now been increased by 17 decibels, with no appreciable change in bandwidth. Lead-network compensation could be employed in a similar manner to increase the bandwidth if this were desirable.

The actual computer program representing $D^*(w)$ is made apparent by deriv-

ing $D^*(z)$, and hence $D^*(z^{-1})$. From equation 33

$$D^*(z^{-1}) = 0.104 \left(\frac{1 - 0.91z^{-1}}{1 - 0.99z^{-1}} \right) \qquad (34)$$

Equation 34 represents a linear computer program requiring a storage for one sampling period of each input and output sample.

Results of Root-Locus and Frequency-Response Techniques Correlated by Means of a Transient Analysis

In order to validate the results of the root-locus and frequency-response techniques, the step-function response of the uncompensated system will be computed by means of a z-transform analysis. For the uncompensated system in which $D^*(z) = 1$, $K_u = 0.63$, and $\alpha = 1.0$, $C^*(z)$ as given by equation 30 can be written as

$$C^*(z) = \frac{0.398z}{z^3 - 2.136z^2 + 1.671z - 0.535} \qquad (35)$$

As is well known[1], if $C^*(z)$ is expanded as a power series in z^{-1}, the coefficient of the z^{-n} term will equal $c(nT)$. Since $C^*(z)$ is a ratio of rational polynomials, the power series is most easily obtained by the simple process of long division, resulting in

$$C^*(z) = 0.398z^{-2} + 0.848z^{-3} + 1.145z^{-4} + 1.233z^{-5} + \cdots$$

The response between sampling instants may be obtained for this system by merely connecting the values at the sampling instants by straight lines since the output of the clamper goes through a process of integration to produce the actual system response. The corresponding linearized system response was calculated for this case and may be written as

$$c_l(t) = 1 - 1.37 \epsilon^{-0.312 t/T} \sin\left(0.681 \frac{t}{T} + 0.816 \right) \qquad (36)$$

The actual system response is plotted in Fig. 10 together with the linearized system response.
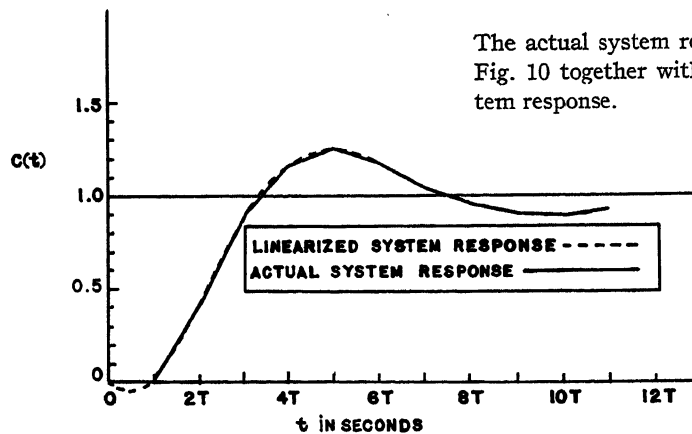
Fig. 10 (left). Step-function response of the uncompensated system

Conclusions

A method of real inversion for the z-transformation has been developed to yield a linearized system response that agrees with the actual system response at the sampling instants. The response so chosen can be related to the poles of the z-transform of the response function in a manner analogous to that for continuous systems. A technique is outlined which allows a ready evaluation of the validity of the use of such a response to characterize the actual system response between sampling instants. The application of root-locus techniques to the design of sampled-data systems is developed where the system stability is selected in terms of the damping ratio of the dominant closed-loop poles that typify the linearized system response.

A bilinear w-transformation is introduced which allows the application of Routh's criterion to sampled-data systems and also permits the use of asymptotic plotting techniques to simplify the construction of the closed-loop frequency locus by use of the Nichols chart. The synthesis of digital computer programs to compensate and improve system performance is easily effected in terms of the desired shaping of the asymptotic frequency locus. Thus the procedures for compensating sampled-data systems are placed on a par with the synthesis of analogue compensation for continuous systems.

A specific sampled-data system is analyzed for a dimensionless sampling period, and a dimensionless velocity constant is specified for a desired degree of stability as determined by root-locus and frequency-response techniques. A linear digital-computer program is synthesized to allow an increase in system gain with no appreciable change in system stability. The foregoing system study is introduced for the purpose of illustration, and should serve as a guide to the designer when utilizing the techniques outlined in this paper.

Appendix I. Derivation of Inversion Integral

If $C(s)$ is the Laplace transform of $c(t)$, then the value of $c(t)$ at the sampling instants $t = mT$ may be obtained from the inversion integral as follows

$$c(mT) = \frac{1}{2\pi j} \int_{\sigma_0 - j\infty}^{\sigma_0 + j\infty} C(s) \epsilon^{smT} ds \qquad (37)$$

When the range of integration is broken up into intervals of length ω_0, where $\omega_t T = 2\pi$, the following result is obtained

$$c(mT) = \frac{1}{2\pi j} \sum_{n=-\infty}^{n=\infty} \int_{\sigma_0+(n-\frac{1}{2})\omega_0 j}^{\sigma_0+(n+\frac{1}{2})\omega_0 j} C(s)\epsilon^{smT} ds$$

(38)

With s replaced by $s+jn\omega_0$

$$\int_{\sigma_0+(n-\frac{1}{2})\omega_0 j}^{\sigma_0+(n+\frac{1}{2})\omega_0 j} C(s)\epsilon^{smT} ds$$

becomes

$$\int_{\sigma_0-\frac{j\omega_0}{2}}^{\sigma_0+\frac{j\omega_0}{2}} C(s+jn\omega_0)\epsilon^{smT} ds$$

(39)

From equations 38 and 39 there results

$$c(mT) = \frac{1}{2\pi j} \sum_{n=-\infty}^{n=\infty} \int_{\sigma_0-\frac{j\omega_0}{2}}^{\sigma_0+\frac{j\omega_0}{2}} C(s+jn\omega_0)\epsilon^{smT} ds$$

Using the standard notation

$$C^*(s) = 1/T \sum_{n=-\infty}^{n=\infty} C(s+jn\omega_0)$$

this becomes

$$c(mT) = \frac{T}{2\pi j} \int_{\sigma_0-\frac{j\omega_0}{2}}^{\sigma_0+\frac{j\omega_0}{2}} C^*(s)\epsilon^{smT} ds$$

(40)

As has been pointed out before, $C^*(s)$ is a rational fraction in ϵ^{sT} in the problems under discussion. Hence the transformation $z=\epsilon^{sT}$, for $-\omega_0/2 < \omega \leqslant \omega_0/2$ gives

$$c(mT) = \frac{1}{2\pi j} \oint_{\Gamma} C^*(z) z^{m-1} dz$$

(41)

where Γ is a circle of radius $\epsilon^{\sigma_0 T}$ with the center at the origin, and $C^*(z)$ is a rational fraction in z.

This transformation maps the semi-infinite strip $-\omega_L/2 < \omega \leqslant \omega_0/2$ into the region $|z| \leqslant \epsilon^{\sigma_0 T}$. Since there are no singularities in $C^*(s)$ on or to the right of $\sigma = \sigma_0$, all singularities of $C^*(z)$ lie inside the circle $|z| = \epsilon^{\sigma_0 T}$. Therefore the path of integration in equation 41 may be any closed contour enclosing all the singular points of $C^*(z)$ when $m \geqslant 1$.

Appendix II. Illustration of the General Method for Solving for C(s) and C*(s)

In the more complicated systems it is not always obvious how to solve for the response or its z-transform. The general method for doing this is illustrated in the following.

Consider the system represented in Fig. 7 for which

$$C = [(C_1 R)^* - (G_1 C)^*]D^* G_2$$

(42)

To solve for C or C^*, it is necessary to eliminate $(G_1 C)^*$ from equation 42. To do this, multiply both sides of equation 42 by G_1 and take the z-transform of both sides. This procedure gives

$$(G_1 C)^* = [(G_1 R)^* - (G_1 C)^*]D^*(G_1 G_2)^*$$

(43)

Table I. Abbreviated Table of Laplace, z, and w-Transforms

f(t)	F(s)	F*(z)	F*(w)
$\delta(t-nT)$	ϵ^{-nTs}	z^{-n}	$\left(\dfrac{1-w}{1+w}\right)^n$
$U(t)$	$\dfrac{1}{s}$	$\dfrac{z}{z-1}$	$\dfrac{1+w}{2w}$
t	$\dfrac{1}{s^2}$	$\dfrac{Tz}{(z-1)^2}$	$\dfrac{T(1-w^2)}{4w^2}$
ϵ^{-bt}	$\dfrac{1}{s+b}$	$\dfrac{z}{z-\epsilon^{-bT}}$	$\dfrac{1}{1+\epsilon^{-bT}}\left(\dfrac{1+w}{\tanh\frac{bT}{2}+w}\right)$
$\sin bt$	$\dfrac{b}{s^2+b^2}$	$\dfrac{z\sin bT}{z^2-2(\cos bT)z+1}$	$\dfrac{\tan\frac{bT}{2}(1-w^2)}{2\left(\tan^2\frac{bT}{2}+w^2\right)}$
$\cos bt$	$\dfrac{s}{s^2+b^2}$	$\dfrac{z(z-\cos bT)}{z^2-(2\cos bT)z+1}$	$\dfrac{(1+w)\left(\tan^2\frac{bT}{2}+w\right)}{2\left(\tan^2\frac{bT}{2}+w^2\right)}$
$\epsilon^{-bt}f(t)$	$F(s+b)$	$F^*(\epsilon^{bT}z)$	$F^*\left(\dfrac{w+\tanh\frac{bT}{2}}{1+w\tanh\frac{bT}{2}}\right)$
$f(t-nT)$	$\epsilon^{-snT}F(s)$	$z^{-n}F^*(z)$	$\left(\dfrac{1-w}{1+w}\right)^n F^*(w)$

Elimination of $(G_1 C)^*$ from equations 42 and 43 gives

$$C = \frac{(G_1 R)^* D^* G_2}{1 + D^*(G_1 G_2)^*}$$

(44)

and hence

$$C^* = \frac{(G_1 R)^* D^* G_2^*}{1 + D^*(G_1 G_2)^*}$$

(45)

Appendix III. Initial- and Final-Value Theorems and Inversion Integral in w

The following theorems are easily obtained from the corresponding theorems in z:

1. Initial value theorem

$$c(0) = \lim_{w \to 1} C^*(w)$$

2. Final value theorem

$$\lim_{m \to \infty} c(mT) = \lim_{w \to 0} [2wC^*(w)]$$

if all poles of $C^*(w)$ are in the left-half plane.

3. Inversion integral

$$c(mT) = \frac{1}{2\pi j} \oint_{\Gamma} C^*(w) \frac{(1+w)^{m-1}}{(1-w)^{m+1}} dw$$

where for $\sigma_0 \neq 0$, Γ is the circle

$$\left|\frac{(w+1)}{(w-1)}\right| = \epsilon^{\sigma_0 T},$$

described in the clockwise direction when $\sigma_0 > 0$ and counterclockwise when $\sigma_0 < 0$. When $\sigma_0 = 0$, the path will be the imaginary axis from $-j\infty$ to $j\infty$. This path may be closed either to the left or to the right when the integrand is a rational fraction in w with the degree of the denominator two greater than the degree of the numerator.

References

1. THE ANALYSIS OF SAMPLED DATA SYSTEMS, J. R. Ragazzini, L. A. Zadeh. *AIEE Transactions,* vol. 71, pt. II, 1952, pp. 225–32.

2. TRANSIENTS IN LINEAR SYSTEMS, VOLUME I. M. F. Gardner, J. D. Barnes. John Wiley & Sons, Inc., New York, N. Y., 1942, Lumped-Constant Systems.

3. A GENERAL THEORY OF SAMPLING SERVO-SYSTEMS, D. F. Lawden. *Proceedings,* Institution of Electrical Engineers, London, England, vol. 98, pt. IV, Oct. 1951, pp. 31–36.

4. SAMPLED-DATA CONTROL SYSTEMS STUDIES THROUGH COMPARISON OF SAMPLING WITH AMPLITUDE MODULATION, W. K. Linvill. *AIEE Transactions,* vol. 70, pt. II, 1951, pp. 1779–82.

5. SAMPLED-DATA PROCESSING TECHNIQUES FOR FEEDBACK CONTROL SYSTEMS, A. R. Bergen, J. R. Ragazzini. *AIEE Transactions,* vol. 73, pt. II, Nov. 1954, pp. 236–47.

———————◆———————

Discussion

Eliahu I. Jury (University of California, Berkeley, Calif.): This paper adds considerable information to the z-transform method and its application for synthesis of sampled-data control systems. The use of the linearized-system response mentioned in this paper for design at the sampling instances is very useful and informative. However, the use of this method for synthesis between sampling instants seems to have certain pitfalls, for one has to check, using equation 28, to find the discrepancy between the final results and the original assumptions on the overshoot and peak time. This procedure requires the evaluation of the quantities in equation 28, which will complicate the design procedures. Furthermore, even at the sampling instants the design information obtained from this method
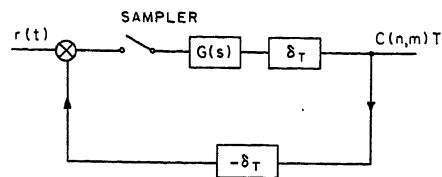
Fig. 11. Sampled-data control system with a fictitious delay to obtain response between sampling instants

Fig. 12 (right). Constant overshoot locus in the z-plane for the following system: $G(s) = \omega_n^2 (1 - \epsilon^{-TS})/s^2(s + 2\omega_n\eta)$, $M_p = C_{max} (n,m)T - 1$, $\omega_n T$, η are variables

is not exact, for it is not always that the peak overshoot and peak time coincide with the sampling instants; indeed, in most cases they do not.

In recent years, few articles have been published in which the z-transform method has been modified to yield information between sampling instants[1-4], but none has discussed any synthesis method utilizing these procedures for the actual design of sampled-data control systems. Extending the method described in reference 4, one can obtain a design procedure which utilizes the extended-z-transform method. To obtain the actual response of a sampled-data control system one may include a fictitious delay in the forward path and a negative one in the feedback path as shown in Fig. 11. The output-input z-transform is given[4] as:

$$\frac{C^*(z,m)}{R^*(z)} = \frac{G^*(z,m)}{1+G^*(z)} \tag{46}$$

where $m = 1 - \delta$, δ is the delay.

If equation 46 is compared with the system without the delay as seen in equation 19, then one can deduce the following observations:

1. The poles of $C^*(z,m)$ are the same as the poles of $C^*(z)$ in equation 19.
2. The zeros of $C^*(z,m)$ in number and location, are different than the zeros of $C^*(z)$.

The fact that only the zeros of $C^*(z,m)$ are different than the corresponding zeros in equation 19 facilitates considerably the design procedures using the z-transform method.

To obtain the actual response $C(n,m)T$, the following contour integral can be used:

$$C(n,m)T = \frac{1}{2\pi j}\int_\Gamma \frac{G^*(z,m)}{1+G^*(z)} XR^*(z)xz^{n-1}dz \tag{47}$$

where Δ is a path of integration in the z-plane that includes all the poles of the integral in equation 47. This integral can be easily evaluated using tables[5] for m is constant in the integration, and to cover the whole response m continually varies from zero to unity, and when m equals unity, one obtains $C(nT)$, the response at the sampling instants. In conclusion, one can notice that the design procedures for a second-order system, using the mentioned modified z-transform, can be readily effected. For a higher order system when dominance holds, the design criterion as to overshoot and peak-time can be easily obtained by varying m from zero to unity to find the most severe condition on these response quantities.

The application of this analysis to the root-locus method mentioned in the paper
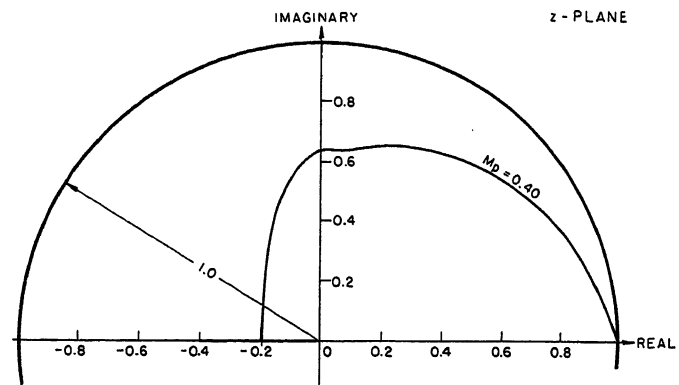
will yield a constant overshoot, M_p, curve which can be used as a design criterion for the second-order system mentioned in this paper. This overshoot is for the actual response between sampling instants, and it is seen to differ from the linearized response. This constant overshoot curve still can be used for a higher order system when dominance holds, and it can be shown[6] that the efpoles of other poles and zeros, besides the dominant pair, on this overshoot can be easily determined. The constant ζ curve plotted in Fig. 8, yields only the complex roots of $1+G^*(z)$ for the linearized response. However, for the system to have the same constant overshoot, the roots can also lie on the negative real axis as seen from Fig. 12. This part of the locus seems not to appear in the linearized-response method of this paper.

The bilinear transformation is a useful tool in the frequency domain design, as indicated clearly in the paper, especially when digital compensation is used. However, if linear network compensation is utilized, the value of the transformation diminishes considerably. Furthermore, this bilinear transformation can be extended to cover the response between sampling instances as well, using the afore-mentioned modified z-transform.

As a point of interest, it can be noted that the method of obtaining the actual response from the linearized response is the same as the impulsive response approach indicated in reference 3.

In conclusion, it should be mentioned that the linearized response method indicated in this paper is a straightforward procedure in the digital computer design of sampled-data control systems with certain limitations indicated, and the consolidation of this method with other methods of design will add substantially to the field of sampled-data control systems.

REFERENCES

1. EXTENSION OF CONVENTIONAL TECHNIQUES TO THE DESIGN OF SAMPLED-DATA SYSTEMS, W. K. Linvill, R. W. Sittler. *IRE Convention Record*, Institute of Radio Engineers, New York, N. Y., pt. I, 1953, pp. 99–104.

2. ANALYSIS OF ERRORS IN SAMPLED-DATA FEEDBACK SYSTEMS, Jack Sklansky, J. R. Ragazzini. *AIEE Transactions*, vol. 74, pt. II, May 1955, pp. 65–71.

3. THE DESIGN OF SAMPLED-DATA FEEDBACK SYSTEMS, Gladwyn V. Lago, John G. Truxall. *Ibid.*, vol. 73, pt. II, Nov. 1954, pp. 247–53.

4. THE PULSE TRANSFER FUNCTION AND ITS APPLICATION TO SAMPLING SERVOSYSTEMS, R. H. Barker. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 99, pt. IV, 1952.

5. ANALYSIS AND SYNTHESIS OF SAMPLED DATA CONTROL SYSTEMS, Eliahu I. Jury. *AIEE Transactions*, vol. 73, pt. I, Sept., 1954, pp. 332–46.

6. SYNTHESIS AND CRITICAL STUDY OF SAMPLED-DATA CONTROL SYSTEMS, E. I. Jury. *Electronics Research Laboratory Research*, University of California, Berkeley, Calif., issue 136, series 60, 1955.

George A. Biernson (Massachusetts Institute of Technology, Cambridge, Mass.): The authors should be strongly congratulated for developing a technique of plotting and analyzing sampled transfer functions which enables the designer to use the same techniques that have proved so useful in dealing with continuous transfer functions. However, they should be criticized for not presenting it better. In spite of the obvious pains that were taken to prepare a good introduction, curiously enough the paper uses only one small paragraph and one equation (equation 22) to present the fundamental contribution of the paper, the z to w transformation. Because of this, I feel that the meaning of the paper is essentially lost to all but the very conscientious readers.

The paragraph of the paper preceding equation 32, states that "it has been shown that the real-frequency plot of $G^*(j\omega) = G^*(jv)$, where $\omega = (2/T) \tan^{-1} v$" but I have been unable to find where this has been shown. After painstaking effort, I finally realized that this proof that was supposed to have been shown, but strangely enough appears to have been lost, represents the key for explaining the basic concept of the paper. Since I feel that the conclusions of the paper are very important, I am taking the liberty to present my own development of that proof, which I finally stumbled upon.

The fundamental transformation of equation 22 in the paper can be expressed as

$$\frac{1}{z} = e^{-sT} = \frac{1-w}{1+w} \tag{48}$$

This transformation has the convenient and well-known property that it works both ways:

$$w = \frac{1-e^{-sT}}{1+e^{-sT}} \tag{49}$$

We are interested in plotting the frequency response of $G^*(s)$, so that we would like to know the values of G^* for $s = j\omega$. Setting s equal to $j\omega$ in equation 49 gives

$$w = \frac{1-e^{-j\omega T}}{1+e^{-j\omega T}} = \frac{1-(\cos \omega T - j \sin \omega T)}{1+(\cos \omega T - j \sin \omega T)} \tag{50}$$

Rationalize the expression by multiplying

numerator and denominator by the conjugate of the denominator, which is $(1+\cos \omega T)+j \sin \omega T$. Multiply the expression out and use the substitution

$$\cos^2 \omega T + \sin^2 \omega T = 1 \qquad (51)$$

This gives the simple expression

$$w = j \frac{\sin \omega T}{1+\cos \omega T} \qquad (52)$$

There is a well-known trigonometric identity that

$$\tan (x/2) = \frac{\sin x}{1+\cos x} \qquad (53)$$

Hence equation 52 can be expressed as

$$w = j \tan (\omega T/2) \qquad (54)$$

Equation 54 shows that for real frequencies, or imaginary values of s, the variable w has, like s, a pure imaginary value. Thus, for real ω frequencies, one can express w in the form

$$w = jv \qquad (55)$$

where v may be considered to represent a pseudo frequency and is related to the actual frequency ω by the simple relation

$$v = \tan (\omega T/2) \qquad (56)$$

The implication of this transformation is that the sampled transfer function $G^*(s)$, which represents an infinite sum when expressed in terms of real ω frequencies, has a convenient closed form when expressed in terms of the pseudo-frequency variable v. Because of this, standard asymptotic techniques can be used for plotting $G^*(j\omega)$ in terms of the pseudo-frequency variable v. This creates a plot which is the same as the actual frequency plot of $G^*(j\omega)$ except that the horizontal frequency scale is distorted, and one has to use the relationship of equation 56, in the form

$$\omega = (2/T) \tan^{-1} v \qquad (57)$$

to determine what a given value of the pseudo-frequency v represents in terms of the actual frequency ω. This distortion of the horizontal frequency scale represents no real problem any more than is the very common scale distortion of plotting the logarithm of the frequency instead of the actual frequency.

This transformation from e^{-sT}, to ω, is very simple (after it has been discovered) and the inherent simplicity may lead many to disregard it. However, the reader should recognize that some of the most important and practical techniques in the field of feedback control are extremely simple. The technique for plotting frequency responses, as log magnitude and phase plots versus a log frequency scale, and the accompanying asymptotic plotting techniques, is almost naive, yet it has proved a tremendous asset. There are, of course, those who refuse to use anything but the polar locus (under the theory that it contains the same information as the logarithmic plots), and there are others who believe that the painfully plotted root locus is the only technique to use. However, if, hypothetically, the field should have to choose between the polar-locus and root-locus techniques, on the one hand, and the logarithmic techniques on the other,

there would be no question as to which would be the more practical and useful choice, and would yield more basic insight into the problems of feedback control.

In order to place the sampled-data techniques into the proper frame of reference, along with the well-known techniques for designing continuous systems, the reader should be reminded that the main dynamic effect of sampling in a feedback control loop is to add phase lag which limits the loop bandwidth. Unfortunately the vast amount of high-powered literature recently published on sampled-data systems has left the average engineer with an inspired feeling but a rather chaotic appreciation of the practical aspects. However, one paper[1] deserves commendation for presenting a good practical analysis and discussion of the effects of sampling. It shows that the action of the clamper, which converts the sampled data back into continuous data, is the source of the main dynamic effect of the sampling process. By considering the clamper alone, it derives frequency-response plots which are good approximations for frequencies below 1/4 of the sampling frequency. The frequency response has essentially a flat magnitude characteristic and a phase lag proportional to frequency, with 45 degrees phase lag occurring at 1/4 of the sampling frequency.

A discussion of this paper criticizes the analysis as being approximate and valid only when the sampling frequency is high enough to permit the replacement of the sampled data device by a continuous one. This is, of course, true but the important point that is not generally recognized is that the gain-crossover frequency, or "bandwidth," of a feedback control loop with sampled data must be less than 1/4 of the sampling frequency to achieve adequate stability.

The manner in which sampling limits the bandwidth can be readily shown by assuming that a sampler and ideal clamper is added to a feedback control loop with the simple open-loop transfer function

$$G_1 = \frac{\omega_c}{s} \qquad (58)$$

where ω_c is the gain-crossover frequency which is the frequency where the magnitude of G_1 passes through unity. The resultant loop transfer function with the ideal clamper is

$$G_2 = \frac{\omega_c}{s} \left(\frac{1-e^{-Ts}}{s} \right) \qquad (59)$$

The ideal clamper holds each pulse until the next one appears. The sampled transfer function G_2^* can be found readily by applying z-transforms to equation 59. This gives

$$G^* = \frac{\omega_c T}{z-1} = \frac{\omega_c T}{e^{Ts}-1} \qquad (60)$$

For real frequencies, this is

$$G_2^*(j\omega) = \frac{\omega_c T}{e^{j\omega T}-1} = \frac{\omega_c T}{(\cos \omega T - 1)+j \sin \omega T} \qquad (61)$$

The magnitude and phase are

$$|G_2^*(j\omega)| = \frac{\omega_c T}{2 \sin (\omega T/2)} \qquad (62)$$

$$\angle G_2^*(j\omega) = -[(\pi/2)+(\omega T/2)] \qquad (63)$$

Comparing this with the original loop transfer function G_1, which was simply ω_c/s, shows that the sampling and clamping increases the magnitude by

$$\left| \frac{G_2^*}{G_1} \right| = \frac{(\omega T/2)}{\sin (\omega T/2)} \qquad (64)$$

and it increases the phase lag as shown by

$$\angle G_2^* = \angle G_1 - (\omega T/2) \qquad (65)$$

Thus the sampling process adds a linearly increasing phase curve, which cannot be compensated for by a lead network because the sampling creates a rising rather than a falling magnitude curve.

From equation 60, the closed-loop transfer function for real frequencies is

$$\frac{G_2^*}{1+G_2^*} = \frac{\omega_c T}{e^{Ts}+\omega_c T-1}$$
$$= \frac{\omega_c T}{\cos \omega T+j \sin \omega T+\omega_c T-1} \qquad (66)$$

Setting the derivative of the magnitude of this function equal to zero, shows that the magnitude peaks where

$$\sin \omega T = 0 \qquad (67)$$

For the lowest frequency above zero, this occurs at $\omega T = \pi$, so that

$$\cos \omega T = -1 \qquad (68)$$

Thus, at the resonant frequency, the closed-loop frequency response of equation 66 reduces to

$$\frac{G_2^*}{1+G_2^*} = -\frac{\omega_c T}{2-\omega_c T} \qquad (69)$$

The magnitude of this is the value of frequency-response peaking M_p.

$$M_p = \frac{\omega_c T}{2-\omega_c T} \qquad (70)$$

Solving for $\omega_c T$ gives

$$\omega_c = \frac{2}{T} \frac{M_p}{1+M_p} \qquad (71)$$

The frequency ω_c is the gain-crossover frequency, or bandwidth, of the unsampled system, and is approximately equal to that of the sampled system. For an M_p of 1.3, this gives for the crossover frequency f_c in cycles per second,

$$f_c = \omega_c/2\pi = f_s/5.5 \qquad (72)$$

where f_s is the sampling frequency. If the gain is increased above this value by only the factor 1.75, the loop is unstable, which indicates that 1/5 of the sampling frequency is about the upper practical limit to the crossover frequency.

Generally, of course, the loop transfer function is nowhere near as simple as has been assumed in the foregoing example. Consequently, the gain-crossover frequency generally is much lower than the limit stated; but, on the other hand, other dynamic factors in the loop cannot allow the gain-crossover frequency to exceed this limit. To achieve reasonable gain in the loop, the loop transfer function must have at least one

net integration for a significant frequency region near gain crossover. If there are more integrations, or lags, this can serve only to lower the bandwidth. Lead networks can be added to counteract the effect of these other lags, but they cannot increase the bandwidth above the value that was achieved in the example, because of the rising magnitude characteristic produced by the sampling. An integral network can increase the low-frequency gain, so as to increase the velocity constant, but it lowers rather than raises the bandwidth.

The sampler that was considered in the example was termed "ideal." However, a sampler characteristic can be achieved that has less phase shift at low frequencies by modifying the slope of the clamper output between samples according to the change in the values of the previous samples; but such a sampler characteristic merely adds the effect of a lead network, and like any other type of lead network it cannot raise the gain-crossover frequency above the value achieved by the example. Similarly, any digitally programmed compensation technique operating on the sampled data has the same basic limitations as continuous compensation techniques.

It should be emphasized that the criteria for bandwidth I have used, the gain-crossover frequency (the frequency where the magnitude of the open-loop is unity), is quite different from the criteria for bandwidth used in the paper, the resonant frequency ω_m. The gain-crossover frequency generally corresponds closely to the frequency of 60 degrees phase lag of the closed loop (C/R) and as shown in reference 2 its value in radians per second is roughly equal to the reciprocal of the step-response rise time, which is the time for the step response to rise to 63 per cent of its final value. In contrast, the resonant frequency is a rather poor criteria for bandwidth, particularly for sampled-data systems, because it defines the closed-loop magnitude curve rather than the phase curve, which may give information as to how much the system is disturbed by high-frequency noise, but give little information as to how much error the system has when following an input. Sampled-data systems often have resonant peaks at 180 degrees phase lag of the closed loop, whereas many systems have resonant peaks at around 45 degrees phase lag. If two systems, one of each type, are set so that the frequencies of 60 degrees closed-loop phase lag are the same, the values of step-response rise-time and the values of maximum error to a ramp input will be essentially the same for the two. However, the resonant-frequency criteria will indicate a much wider "bandwidth" for the sampled-data system. This may mean that the sampled-data system is more noisy, but it is no faster.

The foregoing simplified analysis should serve to illustrate the basic limitation that sampling places on bandwidth, so as to provide a frame of reference whereby the more rigorous techniques can be applied intelligently.

I believe that much more can and should be written to explain in simple, physical terms the practical aspects of sample data. The frequency-response approach of the paper, employing the pseudo-frequency variable v, should prove invaluable in making such a presentation.

REFERENCES

1. AN APPROXIMATE TRANSFER FUNCTION FOR THE ANALYSIS AND DESIGN OF PULSED SERVOS, R. G. Brown, G. J. Murphy. *AIEE Transactions,* vol. 71, 1952 (Jan. 1953 section), pp. 435–40.

2. A SIMPLE METHOD FOR CALCULATING THE TIME RESPONSE OF A SYSTEM TO AN ARBITRARY INPUT, G. A. Biernson. *Ibid.,* vol. 74, pt. II, 1955, pp. 227–45.

R. D. Ormsby (Minneapolis-Honeywell Regulator Company, Minneapolis, Minn.): The authors have presented an interesting new method for the design of sampled-data control systems. The method is clearly outlined, and an example is presented illustrating the design principles for a system which has available a means for digital compensation.

Unfortunately, the authors have failed to show that the additional complication of the w-transformation yields any compensating advantage.

The "three apparent advantages" listed in the paper cannot mean with respect to the root-locus plot in the z-plane. The first and third advantages listed in the paper refer to much less information than is available in a root-locus plot, and the information is considerably more difficult to obtain from the w-transformed system function than from the root locus.

The design of digital compensation to shape the root locus in the z-plane is completely analogous to the design of compensation networks for continuous systems in the s-plane. To illustrate this fact, the root-locus plot for the compensated system used as an example in the paper, is shown in Fig. 13.

Curves may be derived for sampled-data
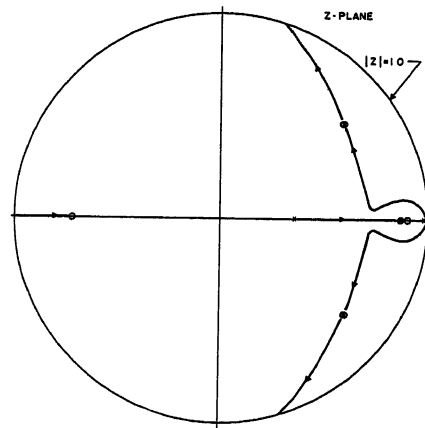
Fig. 13. Root-locus plot of compensated system $\alpha = 1$
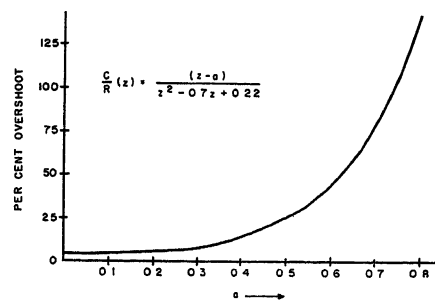
Fig 14. Per cent overshoot in step response versus zero location

systems which relate the transient response to pole-zero configuration in the z-plane (just as is done for a continuous system pole-zero constellation in the s-plane). An example is presented in Fig. 14 where the per cent overshoot in the step response is plotted as a function of the relative position of a pair of conjugate complex poles and a single real zero. The system is shown in Fig. 3 of the paper. Although more profound methods have been used to obtain this information,[1] a very simple procedure may be derived from the principles outlined by G. V. Lago.[2]

REFERENCES

1. THE EFFECT OF POLE AND ZERO LOCATION ON THE TRANSIENT RESPONSE OF SAMPLED-DATA SYSTEMS, E. I. Jury. *AIEE Transactions,* vol. 74, pt. II, 1955, pp. 41–8.

2. ADDITIONS TO z-TRANSFORMATION THEORY FOR SAMPLED-DATA SYSTEMS, G. V. Lago. *Ibid.,* vol. 73, pt. II, 1954, pp. 403–08.

G. W. Johnson, D. P. Lindorff, and C. G. A. Nordling: We appreciate the many interesting comments and suggestions made by the discussers and hope that the paper under discussion will prove to be of interest to many others who are confronted with the design of sampled-data control systems.

Dr. Jury has noted that the linearized continuous response as used in the paper under discussion does not give the exact system response between the sampling instants, and he has cited an interesting and useful methods for deriving the actual response between the sampling instants. However, we feel that the following two points justify the use of the linearized system response as a criterion for design: first, the vast majority of control systems employing sampled data are well characterized as to a margin of stability by the linearized continuous response; second, the several methods that have been proposed in the paper and elsewhere,[1] for evaluating the exact response between sampling instants are primarily useful as methods for analysis, rather than for synthesis. It is believed that the w-transform offers to the designer a synthesis technique in the real frequency domain heretofore restricted to the realm of continuous data systems.

It is of interest to examine Fig. 12 as presented in Dr. Jury's discussion. This figure plots the locus of roots of the second-order system for a constant peak overshoot in response to a step function. With reference to Fig. 8 of the paper it is seen that a contour of constant damping ratio, ζ, corresponds essentially to the locus cited heretofore for bandwidths less than $\omega_0/4$. However, the departure of the constant ζ and m_p contours for larger bandwidths is not a serious discrepancy. It should be remembered that a system design based on the damping ratio of the dominant closed-loop poles is only useful in specifying a range of relative system stability and does not in general uniquely specify the step response overshoot. Thus, for the appropriate systems a 40-per-cent response overshoot might correspond to many different damping ratios usually in the range $0.2 < \zeta < 0.5$. Dr. Jury also states that "for the system to have the same constant overshoot, the roots can also lie on the negative real axis as seen from Fig. 12." However, only a portion of this region, it should be observed, accounts

for a dominant root, namely $-0.4 \leq z \leq -0.2$ which defines a range of damping ratios of $0.28 \leq \zeta \leq 0.46$. Hence, for this region of greatest discrepancy the constant ζ locus is still a reasonable criterion of relative system stability.

The concepts put forth by Dr. Jury are of considerable interest and serve to clarify the ideas embodied in the paper.

Mr. Biernson states that the paper uses only one equation (equation 22) to present the z to w transformation and as such a painstaking effort was required to obtain the relationship between dimensionless frefrequency v and radian frequency ω given in the paper as $\omega = (2/T) \tan^{-1} v$. However, the s to w transformation may be written as

$$w = \frac{e^{sT}-1}{e^{sT}+1} \tag{73}$$

where for $s = j\omega$

$$w = U + jv = \frac{e^{j\omega T}-1}{e^{j\omega T}+1} = j \tan \frac{\omega T}{2} \tag{74}$$

which is a well-known trigonometric identity. Solving for ω in equation 74 results in $\omega = (2/T) \tan^{-1} v$ as stated in the paper. The authors feel that such a straight forward derivation does not require any further clarification.

Mr. Biernson further states that the bandwidth of a feedback control loop with sampled data "must be less than 1/4 of the sampling frequency in order to achieve adequate stability" and as such the describing function simulation for sampling and clamping proposed by Brown and Murphy[2] will serve as an adequate design procedure. We would like to correct this apparent misconception since in many practical situations the bandwidth of a sampled-data control system may be made as high as 1/2 the sampling frequency which is the fundamental bandwidth limitation due to the sampling process. As an illustration the authors will outline the analysis of one such system where the approximate technique of Brown and Murphy erroneously indicate absolute system stability for all values of

loop gain. When an exact analysis is made utilizing the w-transform the system is found to be unstable for an interval of finite loop gains. Consider a unity feedback system employing error sampling, as defined in Fig. 3 of the paper under discussion, where $G(s)$ included a clamper and single dominant time constant as may be typical of a pulsed regulator. Thus $G(s)$ may be expressed as

$$G(s) = \frac{1-e^{-sT}}{s} \cdot \frac{K}{\tau s+1} \tag{75}$$

which readily results in a pulsed loop transfer function as follows:

$$G^*(w) = \frac{K(1-w)}{\tau' w+1} . \tag{76}$$

where $\tau' = (1+e^{-T/\tau})/(1-e^{-T/\tau})$. The closed loop poles are the roots of the equation $1+G^*(w)=0$ which may be written as

$$(\tau'-K)w+(1+K)=0 \tag{77}$$

This equation is seen to have a zero in the right half w-plane for $K > \tau'$ which indicates system instability. However, by contrast, when the describing function approach is used the approximate loop transfer function becomes

$$G_{eq}(j\omega) = G_d(\omega) \frac{K}{j\tau\omega+1} \tag{78}$$

where G_d $(j\omega)$ is the frequency variant describing function for the sampling and clamping process defined as follows[3]:

$$G_d(j\omega) = \left| \frac{\sin \dfrac{\omega T}{2}}{\dfrac{\omega T}{2}} \right| \cdot e^{\frac{-j\omega T}{2}} \tag{79}$$

With reference to equations 78 and 79 it may be seen that G_{eq} $(j\omega)$ has a phase lag of less than 180 degrees for all $\omega \leq \pi/T$. Thus no encirclement of the $-1+j0$ point by the G_{eq} $(j\omega)$ locus is possible and the designer is misled into thinking that the system is stable.

Although the use of approximate tech-

niques may be justified in some instances they must be carefully applied in order to avoid conflicting results. However, we feel that the application of conventional real frequency techniques through use of the w-transform eliminates the need for approximate analysis of sampled-data control systems.

We agree with Mr. Ormsby where he states that "the three apparent advantages listed in the paper cannot mean with respect to the root-locus plot in the z-plane," as the w-transform does not extend the existing techniques for a root locus analysis of sampled-data systems. However, in contrast, as clearly stated in the paper the three advantages listed pertain to synthesis in the real frequency domain. The synthesis of digital compensation to effect the desired shaping of the logarithmic frequency locus cannot readily be accomplished in terms of the z-transform variable. Although the root locus technique is extremely useful for many aspects of system design it does not provide in itself an adequate general technique for system synthesis. In many instances system design must be performed to meet real frequency specifications such as for the case where system excitations are described in terms of their statistical properties.

We feel that a general system design procedure requires an understanding of the correlation among the real frequency domain, complex frequency domain, and time domain techniques and in this regard it would appear that these discussions have served to emphasize and clarify this viewpoint, as relating to sampled-data systems.

REFERENCES

1. DESIGN OF SAMPLED-DATA FEEDBACK SYSTEMS, G. V. Lago, J. G. Truxal. *AIEE Transactions*, vol. 73, pt. II, 1955, pp. 247–53.

2. AN APPROXIMATE TRANSFER FUNCTION FOR THE ANALYSIS AND DESIGN OF PULSED SERVOS, R. G. Brown, G. J. Murphy. *Ibid.*, vol. 71, 1952, pp. 435–40.

3. A FREQUENCY RESPONSE ANALYSIS OF SAMPLED-DATA CONTROL SYSTEMS, G. W. Johnson. *Master of Science Thesis*, University of Connecticut, Storrs, Conn., 1954.

Modernizing Service Controls on Rapid Transit

GEORGE KRAMBLES
NONMEMBER AIEE

EVER-increasing street traffic congestion has directed renewed interest in rapid transit railroads as the most effective means for bringing large numbers of people swiftly across the broad expanse of modern urban areas. However, the staggering capital outlays required for this kind of facility demand that automation be applied to keep the price of service within range of the market. Fortunately, these same mechanical and

electrical aids produce an improved control over the quality of the service offered.

Line supervision is a collective term for new methods developed by the Chicago Transit Authority (CTA) to provide evenly spaced, equally loaded rapid transit trains to maintain service according to schedule, to restore service following delays, and to increase system capacity. About $125,000 has been invested in it, from which there have been obtained

savings in rolling stock, manpower, and operating costs which easily return the entire investment in 1 year. For example it was estimated that at least 28 to 30 additional cars, worth nearly $1,500,000 at today's prices, are saved through the ability to fit capacity more closely to de-

mand and by reductions in running time and recovery time allowed at terminals to protect return service against delays. Brief consideration of the background of the problem may help to bring out this point.

Background of the Problem

Rapid transit schedules are designed on the basis of the number of passengers requiring service in each 15-minute period of the day and night. Headways between trains and number of cars per train are varied to obtain a balance between economy and passenger convenience that yields the maximum over-all benefit. Although service demands change throughout the day, the transition is gradual and there are long periods of practically constant headway.

Running time is allotted on the basis of average performance, but there are times when a train does not move as fast as it should. Delays are caused by bad weather, variations in car makeup, heavier-than-average traffic, mechanical difficulties, slow crews, or slow passengers. Chicago's rapid transit lines are particularly vulnerable to delays resulting from conflict with other trains at junctions, practically all of which involve track crossings at the same grade. Other times a train can get ahead of time because of

lighter-than-average traffic, extrafast train, extrafast crew, etc., being in its favor.

In peak hours, on CTA's principal rapid transit line, trains are scheduled as close as 2 minutes apart and each picks up about 800 passengers. Capacity of a crowded train is 1,000, so that a deviation in headway exceeding 25 per cent results in overloading that slows a train up, thus aggravating the gap. A train only 1 minute or 2 ahead of time can miss most of its load, leaving the following train a nearly impossible job. Likewise, a train only 1 or 2 minutes late will have much more than its share of work to do if trains ahead have gone through on time.

Uneven service is objectionable to the passenger because he may have to wait longer for his train, he may miss connections at transfer points, and he may find the train following the gap too crowded for comfort. Uneven service is troublesome to the employee because he finds it difficult to work an overloaded train, he finds train congestion at terminals, and he is rushed by shorter time at the end of the line before beginning his return trip. Accumulation of passengers in peak hours is so rapid as to present potential for complete breakdown of service in the event of a delay of only 5 or 10 minutes. Bunching of trains may lead to power failure by concentrating too much rolling stock for the available substation capacity.

Line supervision deals with small service irregularities by controlling the spacing of trains, and with major service interruptions by centrally directed, quickly executed reroutes or turnbacks. It includes modern tools for dispatching trains, for recording train movements, and for direct communication to operating personnel and passengers.

Supervisory Equipment

AUTOMATIC TRAIN DISPATCHING

Line supervision requires accurate time, furnished by a high-quality pendulum master clock. This clock closes a control circuit momentarily every 15 seconds to operate a set of schedule controllers. Each controller determines, by means of a rolled paper tape, when trains are to leave the associated terminal or time point. Holes have been punched into this tape at the exact time that each train is to leave throughout the 24-hour day, with different tapes for weekday, Saturday, Sunday and, in some cases Monday operation. Fig. 1 shows a schedule tape being adjusted in a controller.

Each time a punched hole in a tape

comes into register with the probing finger of the schedule controller, an associated automatic train dispatcher puts on the starting lights at the distant station platform through a low-voltage d-c control sent over the telephone cable. As soon as conditions permit acceptance of the starting signals, the conductor closes the doors and the motorman starts his train. As it passes over an insulated section of track, another circuit is operated over telephone lines to the automatic train dispatcher putting out the starting lights. Fig. 2 is a simplified schematic diagram of the basic automatic train-dispatching circuitry.

Three manual control keys above the recorders at the office permit introducing an extra train, annulling a train, annulling the entire schedule, and advancing or retarding the schedule tape. This gives the flexibility needed for dealing with various routine problems as well as with major disruptions. For Congress terminal in the Milwaukee subway, the automatic train dispatcher not only provides the starting signal impulse but also initiates the operation of an unattended automatic interlocking plant to provide a track line-up for the proper trains at starting time.

GRAPHIC TRAIN RECORDS

A continuous graphic record is made of the operation of automatic train dispatchers and of the passage of trains at sufficient intermediate points to provide adequate remote supervision. Fig. 3 shows a sample section of a graphic train chart and, in map form, the locations where supervisory indication or control is taken. Six operation recorders of 20-pen capacity each are arranged in one desk at the central office. An auxiliary desk with three recorders that can duplicate any of the main recorders gives an extra operating position for peak periods.

At the instant a starting signal is given, a record is made on the recorder charts. Similarly, when the train actually leaves, a record is made and as it proceeds along the line its progress is recorded by successive pens so that a series of chart pips from lower left to upper right is made by each train. Fig. 4 is a closeup of one of the recorders.

At certain junctions the graphic indication is modified so that a single recorder pen shows not only the passage of a train but also its route. This is done by pulsing the pen magnet feed if the indication is for the secondary route so that the resultant chart pip is inked in solid. Evenly spaced service shows on the recorder charts as evenly spaced pips.

LEGEND

○ PILOT LAMP OF COLOR INDICATED BY INITIAL
—||—CONTACT CLOSES WHEN DEVICE IS ENERGIZED
—||*—CLOSES ONLY WHEN CONTACT FINGER PENETRATES SCHEDULE HOLE PUNCHED IN TAPE
—//—CONTACT CLOSES WHEN DEVICE IS DEENERGIZED
—• •— Normally open LEVER KEY
—•▮•— Normally closed
—⊏—CAPACITOR
—⌐—CONTROL SWITCH
—▢—RELAY COIL
PARTS of ADD-SUBTRACT STEPPER RELAY:
  S="Start" Notching Magnet & interlock contacts
  L="Leave" " " " " "
—⊶= Contact disks
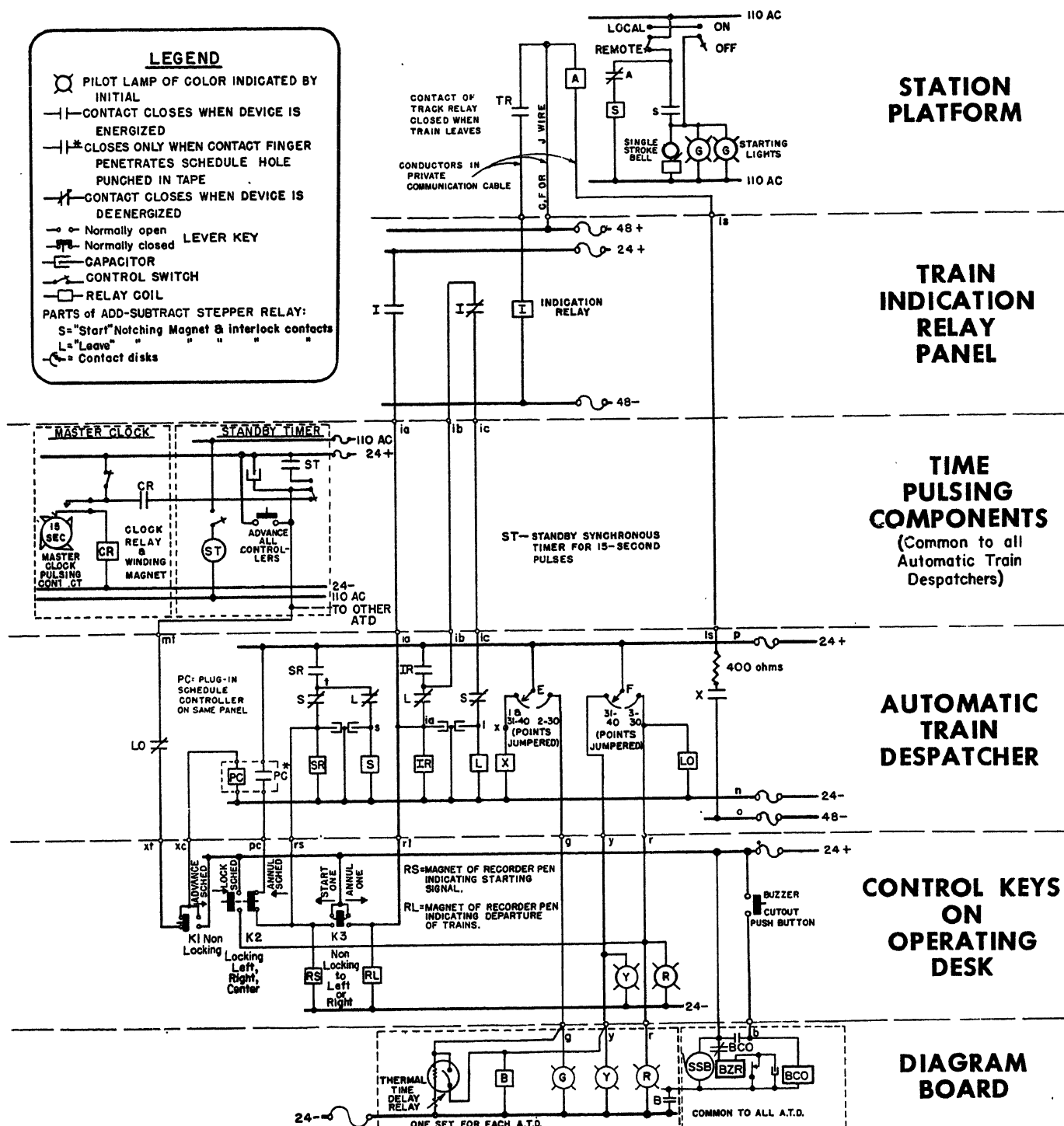
Fig. 2. Automatic train-dispatching simplified schematic diagram

Conversely, irregular spacing resulting from late or early trains is clearly visible on the charts as irregularly spaced pips. Although the system is primarily intended for providing immediate correction to road difficulties, the charts provide much valuable reference material for service improvement studies, complaint analysis, and other by-product uses.

TRAIN INTERVAL TIMERS

Since there are about 90 check points spread over the recorders, it would re-quire extreme concentration for the operator (only one is on duty during lighter hours) to be sure of observing any deviation that might occur. Therefore, train-interval timers are used to provide automatic supervision of key check points.

The train-interval timer is an adjustable process timer connected with the train-passing indication circuits in such a way that the timer mechanism is reset and its motor restarted each time a train passes the associated check point. If successive trains fail to pass within the timer setting, the unit energizes an annunciator light upon reaching its limit. Fig. 5 shows one type of timer unit used.

DIAGRAM BOARD AND OTHER SUPERVISORY LIGHTS

On a wall map are shown all main line tracks, crossings, crossovers, and turnouts with colored lamps set at positions corresponding to the field check points. Illumination of the colored lights describes the nature of deviation which the automatic supervision has detected:

• TRAIN DESPATCHING POINTS

(Automatically controlled by line supervisors' equipment)

• TRAIN DETECTOR LOCATIONS

(Train passing automatically recorded in line supervisors' office)

• GAP TIMER LOCATIONS

(Delays automatically recorded in line supervisors' office)

Fig. 3. Typical section of train graph and map showing locations where indication or control is taken

LEGEND

⊖ Train Detector

⊘ Train Detector & Gap Timer

◄ Train Detector & Terminal Automatic Train Despatcher

▷ Train Detector & Midline Automatic Train Despatcher

⌐⌐⌐ Recommended Additions
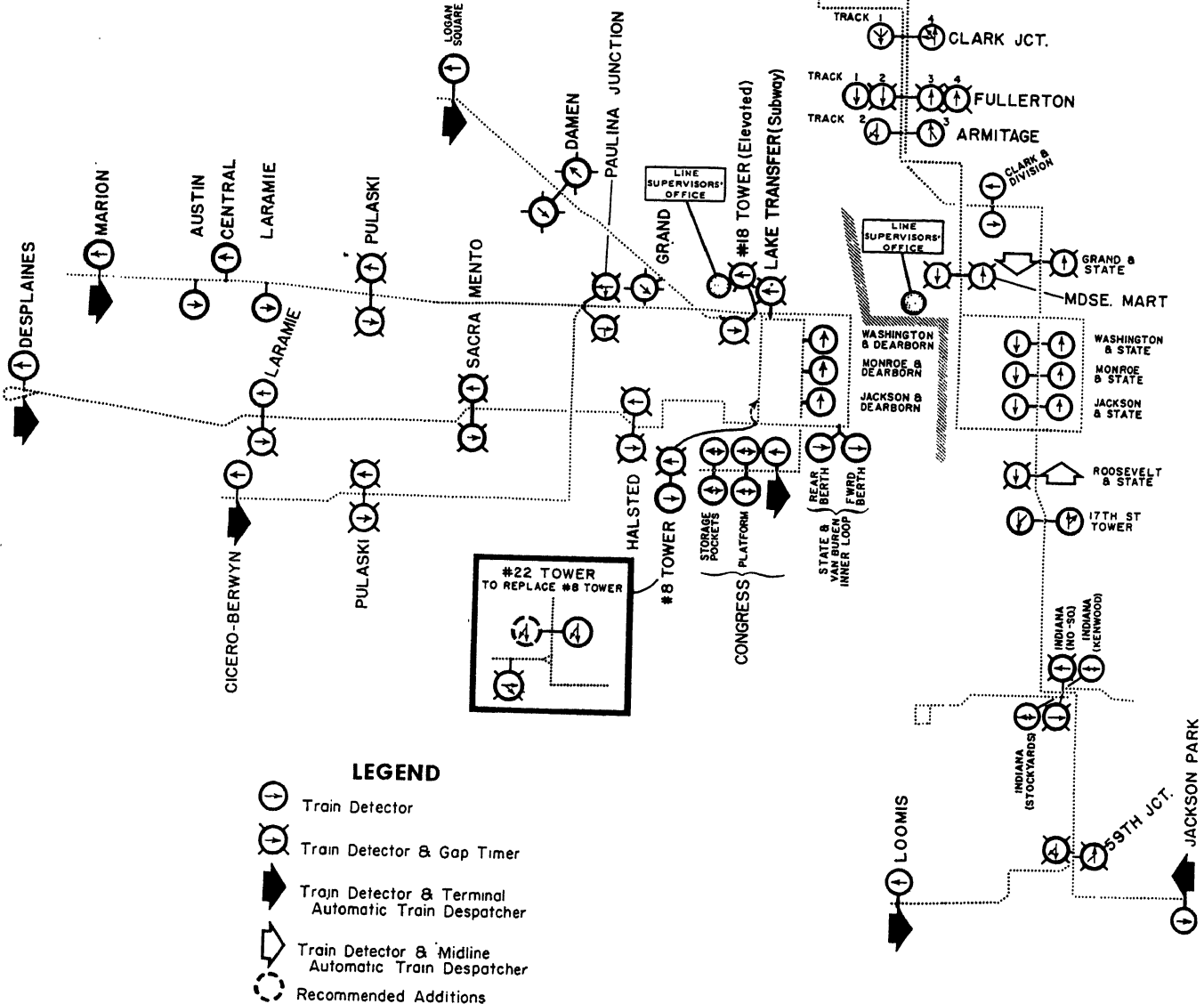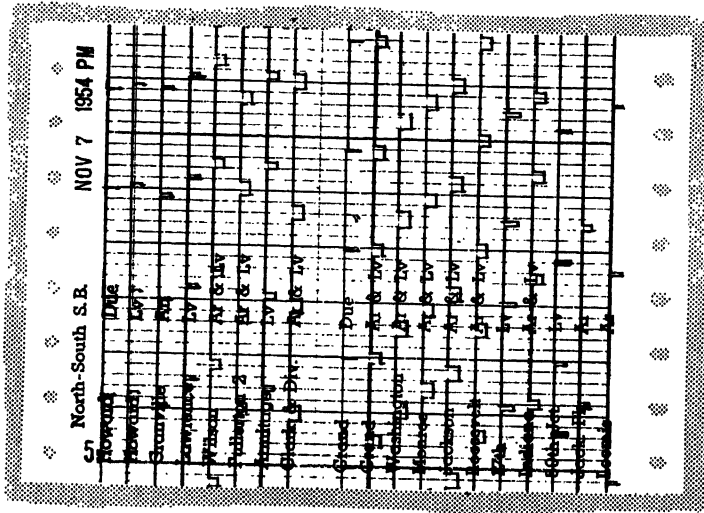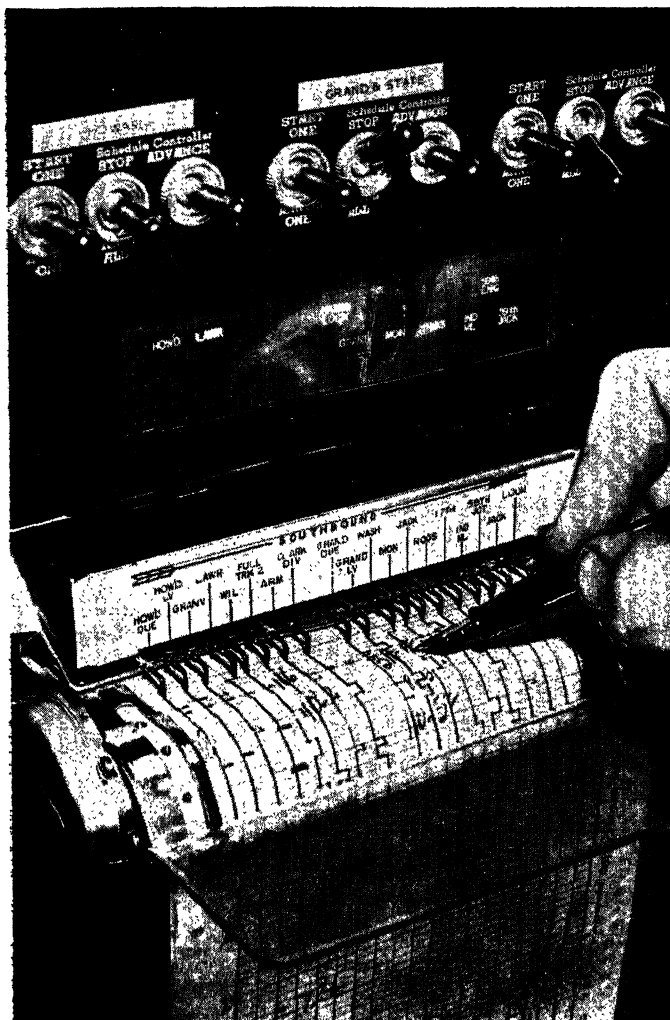
Fig. 4 (left). Close-up of train graph recorder and automatic dispatching controls

Fig. 6 (above). Colored lights on the diagram board indicate the status of train dispatching at each terminal and locations where delays are developing

Fig. 5. One type of process timer adapted for checking intervals between trains

Steady green, starting lights are on.

Flashing green, outbound track lineup is being called for at an automatic terminal preparatory to a train departure.

Yellow, train has left automatic dispatching point without central office authority.

Green and yellow, train due to leave is more than 1 minute late.

Steady red, schedule controller is stopped.

Flashing red, interval between trains is excessive.

A glance at the diagram board shows the location where delays may be developing, but it is necessary to refer to the recorder charts to determine the nature and extent of disruption, the number of trains available ahead of or behind the gap in service, and other factors that determine the proper course of action. There is a duplicate switchboard lamp directly above the recorder pen for the associated check point. Thus, when a delay occurs, the diagram board lamp indicates the geographic location, track, and direction of the gap, and the recorder annunciator lamp spots the chart line which has recorded train operation at that check point. Fig. 6 shows the general arrangement of the diagram board as seen by the line supervisors.

TRAIN DETECTOR DEVICES

In many cases, track circuits which existed for the operation of railway signals were utilized to give line supervision indications at little cost. At check points in nonsignalled territory an inexpensive track circuit was installed, using only one rail length, a cheap annunciator transformer, and a telephone-type relay.

TELEPHONE CABLE CIRCUITS

Connections between field and office are provided by the railway telephone cable. This cable reached all points that were to be checked with the installation of relatively short drop wires. By a fortunate coincidence, cable capacity became available because of the closing of a number of uneconomical local passenger stations, each of which had had a phone. Some research is now being done on methods of multiplexing indications to increase the amount of information which can be transmitted over the existing cable plant.

Communications Equipment

The rapid transit lines of CTA have been covered by a modern private-branch-exchange phone system for many years. However, in cases of service deviations, the busy lines, wrong or forgotten numbers, and delays in response of conventional telephones were such a handicap that it was usually not practical to take action for minor delays and it often was impossible to co-ordinate action for major delays. A new private communication system was therefore developed for line supervision.

LINE SUPERVISION INTERCOM SYSTEM

Fig. 7 is an organization diagram of the special intercom system. At present, it links the central office with 34 key field locations of which 27 are part-time or full-time field supervisor or towerman posts. These locations are distributed over three party-line channels radiating from the main office. Each party-line channel consists of two pairs of line wires, one carrying messages toward and the
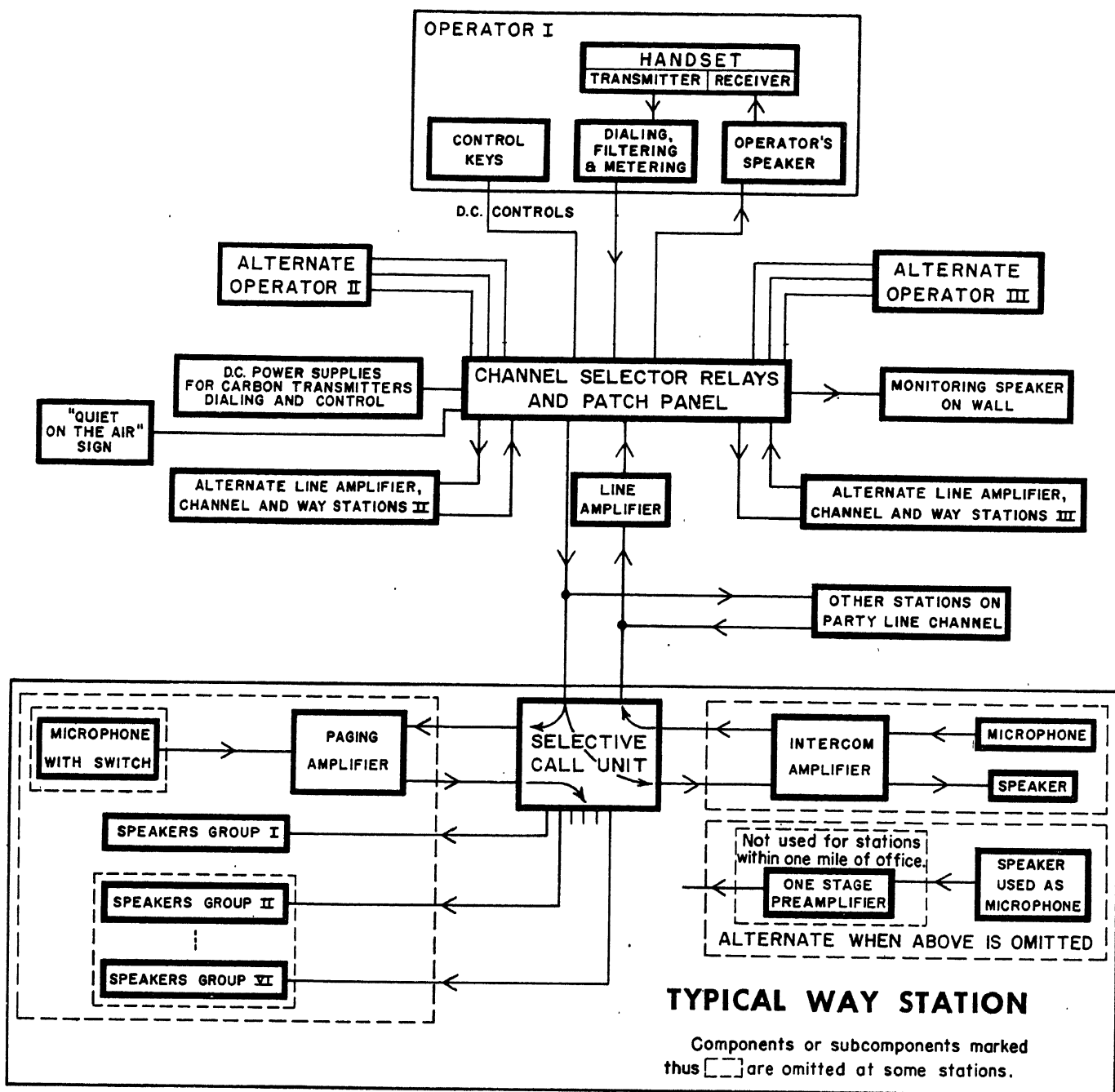
Fig. 7 (above). Simplified organization diagram of line supervisor's private communication system

Fig. 8 (right). Paging loudspeakers and dispatching starting lights are located overhead along edge of platform

other carrying messages away from the central office.

Each station is connected to its channel through a selective call unit which permits central to contact individually any desired station upon dialing a 2- or 3-digit code. Way stations can initiate a call by merely pressing a momentary contact key. Stations not called are not disturbed with messages which do not concern them. Certain stations are so coded that they may be called in a group for simultaneous reception of a main office directive.

The 2-pair channel arrangement automatically makes the central line supervisor one party of all calls and prevents conversation between 2-way stations. This reduces loading of the system and directs all calls with the desired chain of command. The channel arrangement also permits loudspeaker reception without manipulation of talk-listen switches. This saves precious seconds and reduces the possibility of missing parts of messages. Loudspeaker reception permits voice calling and eliminates the delays and expense involved in ringing systems.

Fig. 9.   Intercom and platform paging speaker
locations on the system

COMMUNICATIONS

● Line supervisors' private intercommunication
   system stations in towers, booths, or offices.

— Present and recommended platform paging
   systems operated by line or local supervisor.

LINE
SUPERVISORS'
OFFICE

LINDEN

HOWARD

KIMBALL

BRYN MAWR

LAWRENCE

WILSON

CLARK

LOGAN SQUARE

FULLERTON

ARMITAGE

LINE
SUPERVISORS'
OFFICE

MARION

PULASKI

GRAND & STATE

MERCHANDISE MART

#18 TOWER

RANDOLPH
& WELLS

CLARK & LAKE

WASHINGTON
MONROE   & STATE
JACKSON   & DEARBORN

DESPLAINES

LARAMIE

#12 TOWER

STATE & VAN BUREN

CICERO-BERWYN

PULASKI

#8 TOWER
(TO BE REPLACED
BY #22 TOWER)

ROOSEVELT & STATE

CONGRESS

17th ST.

OUTLYING
SUPERVISOR'S
LOCATION

INDIANA

LOOMIS

61st ST.

PLATFORM COMMUNICATIONS
(Information to Crews
and to the public)

JACKSON PARK

Fig. 10. Automatic operation of junction switch is initiated by car-borne hoop-shaped coil passing wayside transceiver at Lake-Paulina junction

Whenever the line supervisor is using a channel, a red pilot light is displayed at each station on that channel so that a station desiring to initiate a call will not unintentionally interrupt a conversation. However, if necessary, any station may break in to reach the central office with an urgent message.

Amplifiers feeding intercom loud-speakers have automatic limiter action to insure a readable signal through normal variations of voice timbre and volume without readjustment of controls. This feature also permits monitoring the way station and in most cases makes it unnecessary to be especially close to the way station microphone to be clearly heard.

PAGING SYSTEM ON STATION PLATFORMS

Some 42 train platforms at 30 stations are equipped with paging loudspeakers. These are used to reach train crews and the public with announcements. Each such platform layout includes at least one speaker opposite the berthing position of every car. Fig. 8 shows the location of speakers at a train in a typical installation. The speakers are operated at a relatively low volume level, which provides a clear signal along the entire platform and inside each car without producing excessive sound spillage onto abutting property. Automatic limiter amplifiers are used for platform work also to minimize the variations in voice quality without requiring continual readjustment of controls.

At most of the stations the platform speakers can be operated by a local field field supervisor working from a booth

Fig. 11. Each 6-car train represents a capital outlay of more than $300,-000; keeping it in continuous smooth production of transit service is the responsibility entrusted to modern electronic controls

within view of the area covered. At all stations the line supervisor can make announcements directly onto these platforms when no local supervisor is on duty. Fig. 9 shows, in map form, the locations which are equipped with the new communication equipment.

APPLICATION OF SUPERVISORY DEVICES

By knowing that there is a gap in service and just how much it amounts to, the line supervisor takes action to protect or advance the delayed train. Assistance is rendered and correction of the unevenness is accomplished by:

Respacing—holding leaders of a delayed train at points ahead of the delay, to pick up additional load.

Running—ordering a seriously delayed train to pass certain stations without stopping and thus recover all or part of its delay. This is usually done when a delay approaches the scheduled headway so that a second train is immediately following. Runs are required if delays occur during peak hours when trains are loaded nearly to capacity. They are often needed to get defective trains out of the way.

Rerouting—sending trains to stations or destinations other than scheduled to provide alternate service in the event of major delay or total blockade of a track or line.

In cases where car equipment trouble is suspected, the line supervisor sends the

first available supervisor to meet the defective train to assist crews in clearing trouble or limping into terminal. When he has sufficient information he also arranges with terminals to replace trains and, if necessary, to provide substitute crews so that "kickback" delays on return trips can be minimized.

Respacing is done accurately when a delay is spotted on the charts by noting the total time lost by the delayed train and dividing this among the number of leaders known to be between the delay and a control point ahead of the delay. The first of these leaders is held a small amount of time, the next a little more and so on until the delayed train appears on an even, though retarded interval. This technique is only effective when the spacing can be accomplished without overloading, the trains held.

At Grand and Roosevelt in the State Street subway, complete automatic train-dispatcher installations have been made, even though these are not terminals. In case of delay in the lines back of these points, the respacing operation can be carried on through manipulation of the automatic train dispatcher controls by the line supervisor himself. In case of delay along the line which would make a train so tardy arriving at a terminal as to be unable to begin its return tip on time,

the respacing technique is applied to the service starting at the remote terminal. This operation can be carried out accurately without assistance of any local personnel and within minutes the charts provide proof of the effectiveness of the spacing job done.

APPLICATION OF COMMUNICATIONS

Running trains express past scheduled station stops, even though it is the only practical way to break up long delays, has been a dreaded therapy because of the possibility of carrying passengers by their destinations. Before the advant of platform loudspeakers it was seldom done when there were more than a few passengers on a train, but now it can be successfully accomplished under almost any conditions if circumstances warrant. The procedure begins with an announcement before the train concerned arrives, so that waiting passengers will not attempt to board a train which cannot serve them. Stations with paging speakers also have track indicators. Thus, by watching the train graph, the line supervisor sees when a train has arrived. The monitoring feature of his intercom enables him to listen to background noises at the station and thereby to estimate when the train has stopped. The announcement is then repeated until it seems reasonable that it has been heard. New car equipment added in the last 3 years is equipped with interior loud speakers over which the conductor of a train makes regular station announcements or repeats special information he hears on the platform loudspeakers.

Other New Developments in Automatic Transit Controls

AUTOMATIC JUNCTION CONTROL

Some rapid transit routes share common trackage over the inner or downtown portion of their routes and then diverge to cover wider areas in the outlying parts of the city. Given a general knowledge of railway signalling methods, it is understandable that control and protection of junction operation can be made inherent in an all-relay interlocking system to prevent conflicting inbound and outbound moves when these tracks cannot be separated in elevation. In the Lake-Paulina junction between CTA's Lake and Douglas routes, all the converging as well as diverging, conflicting, and nonconflicting moves are made fully automatic and on a first-come-first-served basis. New in this installation are the identification of divergent trains and the sequential identifica-

Table I. Chicago Transit Authority Line Supervision Data Sheet

Rapid Transit System	In Service 1-1-55	To Be in Service 12-1-55
Central Line Supervision Office Coverage:		
Miles of rapid transit route supervised	138	138
Locations from which trains are started by automatic train dispatching:		
End-of-line	9	9
Selective: one of three routes using a certain terminal, obtained by towerman co-operation	1	1
Midline	2	2
Total automatic train dispatching	12	12
Distance to most remote terminal, miles	14	14
20-pen operation recorders used	9	9
Indications recorded on charts:		
Pens indicating trains standing or passing	90	93
Pens indicating movable bridges	3	3
Pens indicating train starting signals	12	12
Above pens on which a second item of information is obtained by pulsing	20	25
Train interval timers used to indicate gaps in service at midline points	35	35
Supervisors' locations connected to line supervisor's loudspeaking intercom system	27	27
Stations equipped with platform loudspeakers	27	30
Platforms equipped with loudspeakers	36	42
Approximate number of platform loudspeakers used	360	442
Distance to most remote intercom and station speaker installation, miles	14	14
Number of amplifiers on continuous 24-hour duty	54	56
Operators on duty weekday rush periods	3	3
Operators on duty Saturdays and base periods	2	2
Operators on duty Sundays and midnight periods	1	1
Outlying Line Supervision Coverage:		
Recorder locations in addition to Central Office	2	2
20-pen operation recorders used	2	2
Indications recorded on these charts	13	18

Surface System	Present	Proposed
Gas bus routes supervised	1	1
Trolley bus routes supervised	1	3
Mileage of routes supervised:		
Motor bus	6	6
Trolley bus	16	43
Indications recorded on chart of buses passing:		
Motor bus	1	6
Trolley bus	4	8
20-pen recorders used	1*	1
Bus interval timers used		10
Schedule controllers used for comparative purposes		2
200-kc transmitters used on motor buses for activating wayside indicating receivers	2	20
Wayside receivers used for motor bus indications	1	6

* Temporarily being recorded on present rapid transit recorders.

tion storage system.

Identification of all convergent trains is obtained from the track circuits approaching the junction. Identification of divergent trains is established from radio-frequency circuit actuated only when a car-borne tuned coil passes the wayside transmitter receiver. Fig. 10 shows the identification installation.

Above the relay case are positioned the transmitting and receiving antenna coils. An 80-kc signal is continuously beamed across the track at this point and, when intercepted by a tuned doughnut-shaped car-borne coil, resonance is produced. When amplified, this actuates momentarily a relay which encodes the track circuit information produced by that train, identifying it for the Douglas route. An approaching train not carrying the coil actuates the same track circuit but, in the absence of the encoding, it is accepted for the Lake route. This system was developed by the Union Switch and Signal Company. Other systems of train identification are under study.

LINE SUPERVISION OF SURFACE VEHICLES

Automatic supervisory systems for surface transit operation are a desperate need. Street car companies pioneered in such devices 40 years ago using trolley-actuated recorders, but the high cost of line wires needed to bring supervision to a central point stymied real development. Conversion from street cars to motor buses eliminated the physical contact as a means of vehicle detection. As street traffic problems become greater and the economic facts of transit life press for fewer men and vehicles with less supervisory manpower, the demand for improved control becomes more imperative.

The CTA has had research on this problem in progress for several years. A lightweight, compact 200-kc transistorized transmitter has been developed which offers promise as one type of link between the bus and the fixed check point. Line supervision will be extended to surface operations as rapidly as techniques can be found.

Table I shows a line supervision data

sheet for both the rapid transit system and the surface system.

Conclusion

With the control system which the CTA is now operating, central line supervisors team up with relatively few field supervisors to keep rapid transit trains rolling. It has resulted in savings in operating costs through closer scheduling with less reserve capacity retained as insurance against overloading caused by service deviations. It has released manpower from train-starting and checking functions formerly scattered over the line.

The fact that a record chart is automatically made impels the supervisor to do his best in dealing with each delay as his work is put on record. It also helps him prove the soundness of his technique, which encourages even better future work. Likewise, the train crews who seldom see the details of the line supervision system respond to the knowledge that an accurate record of their work is made and that they will receive assistance promptly when troubled conditions arise. Announcements made calmly and clearly over loudspeakers elicit passenger co-operation rather than opposition even when the announcement conveys news that will inconvenience some people. Messages delivered the old way by trainmen or platform men invariably led to arguments and further delay.

Optimum Switching Criteria for Higher Order Contactor Servo With Interrupted Circuits

S. S. L. CHANG
MEMBER AIEE

IN A contactor servo, the controlling elements determine only the instants of time at which the contactor operates. In between these moments the controlled system is left alone to follow its own equations of motion and is not in any way affected by the error and the time derivatives of the error. This is also true for continuous control systems with saturable elements in response to a large reference input. For optimum operation of a second-order system of this nature, Hopkin[1] and McDonald[2] showed that the switching criterion can be expressed as a nonlinear relationship between the error and its first time derivative. Their analyses were in terms of the phase-plane technique and the optimum switching boundary was shown to be a trajectory on the phase plane. Bogner and Kazda[3] generalized the phase-plane criteria for second-order contactor servos[1,2] to a phase-space criteria for higher order contactor servos. An implicit but basic assumption of their analysis is that in an n-order contactor servo the error and its derivatives up to the $n-1$ order are continuous at the moment of switching. What does this assumption imply physically?

This assumption would be universally true in a second-order position control system; for as the torque is limited to its saturation value, the error and its first derivative must be continuous at all times. However, in higher order contactor servos, continuity of the second or higher derivative of the error implies uninterrupted circuits during switching, i.e., the circuits remain closed while at the moment of switching the applied electromotive force jumps from a definite value of zero or reverses in value. While switching this way is not only possible but also desirable from the point of view of contactor life, it is by no means the only possibility. An alternative is for the contactor to open the circuit completely at the moment of switching. Because of the large momentary counter voltage at the instant of open-circuit switching, the values of the higher

derivatives drop to zero in a negligible time; hence zero initial values for these derivatives may be assumed as regards subsequent operation.

The optimum open-circuit switching criteria for a high-order contactor servo is investigated in this paper by an extension of the phase-space method. It is found that for contactor servos of the same order the ones with open-circuit switching require less time and less steps to reach a steady state. The computer, which is the counterpart of the controller in a linear servo, is considerably simpler in construction in contactor servos with open-circuit switching than that with closed-circuit switching.

Nomenclature

$P_r(\ )$; $Q_r(\ )$; $F(\ )$ = functions of the parameters or variables in parentheses
$\delta = \pm 1$ represents contactor positions
τ_ϕ = field time constant = L/R
t = time after each switching
J = inertia of motor and load
T_m = motor saturation torque
θ_i = servomechanism input variable
θ_0 = servomechanism output variable
N = order of contactor servo
M = number of nonvanishing initial values of the error and its derivatives
$E = \theta_i - \theta_0$
i = motor field current
r = running index

An Illustrative Example

The foregoing discussion can be exemplified by solution of a typical problem using the circuit of the third-order contactor servo given as an example by Bogner and Kazda.[3] Referring to Fig. 1, the field current i will remain unchanged at the instant of switching only if the computer keeps the motor field circuit closed while reversing or removing the field voltage V. If the computer is to cause the operation of an ordinary contactor with reverse and neutral positions, i must cease to flow at the instant the circuit is opened. While the energy stored in the field inductor has to be disposed of, generally

this can be accomplished with an interval of time equal to approximately 1 or 2 per cent of the field time constant by various antiarcing means. For practical purposes, therefore, i can be considered as discontinuous and starting from zero at the beginning of each switching. The interruption of i also makes a difference in the final balanced state. With closed-circuit switching, the final state must be such that the error, error velocity, and torque are zero simultaneously. With open-circuit switching, it is only necessary that the error and velocity be zero. The torque will be reduced to zero the moment the switch is opened anyhow.

The equation describing the system in the intervals between switchings is

$$\tau_\phi \frac{d^3E}{dt^3} + \frac{d^2E}{dt^2} = \frac{T_m}{J}\delta \tag{1}$$

The derivation of equation 1 is straightforward. It is given in detail in Appendix I.

Let the phase-plane variables U_1 and U_2 be defined by $U_1 = E$, and $U_2 = (dE)/(dt)$. Continuity of the phase trajectories implies the preservation of position and velocity at the moment of switching. The open-circuit switching condition is

$$\frac{dU_2}{dt} = 0 \text{ at } t = 0 \tag{2}$$

Under the condition of equation 2, equation 1 reduces to

$$\tau_\phi \frac{d^2U_2}{dt^2} + \frac{dU_2}{dt} = \frac{T_m}{J}\delta \tag{3}$$

Solving equations 2 and 3 yields

$$\frac{dU_2}{dt} = \frac{\delta T_m}{J}\left(1 - \epsilon^{-\frac{t}{\tau_\phi}}\right) \tag{4}$$

Integrating equation 4 gives

$$U_2 = \frac{\delta T_m r_\phi}{J}\left(\frac{t}{\tau_\phi} - 1 + \epsilon^{-\frac{t}{\tau_\phi}}\right) + U_{20} \tag{5}$$

$$U_1 = \frac{\delta T_m \tau_\phi^2}{J}\left(\frac{t^2}{2\tau_\phi^2} + 1 - \frac{t}{\tau_\phi} - \epsilon^{-\frac{t}{\tau_\phi}}\right) + U_{20}t + U_{10} \tag{6}$$

Plot of U_2 and U_1 in equations 5 and 6 as the ordinate and abscissa respectively yields a phase trajectory for each pair of values of U_{10} and U_{20}. t is a parameter with the physical meaning of time which a point representing a state of the system
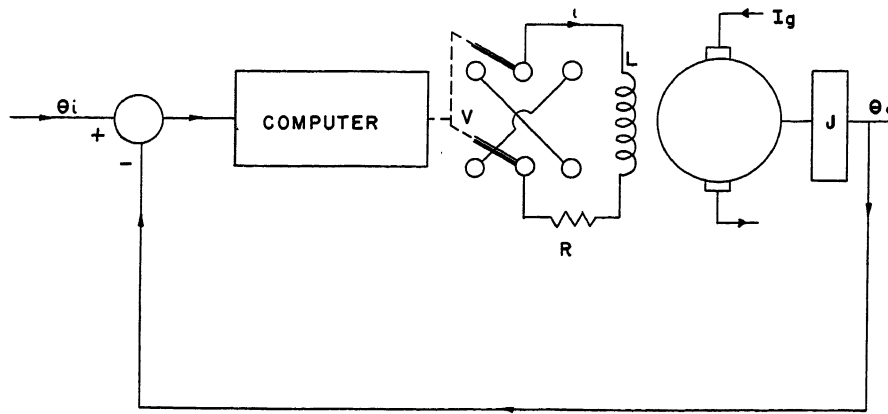
Fig. 1. Block diagram of third-order contactor servo

takes to traverse from the initial point on the phase plane to the point (U_1, U_2) in question. As some of the trajectories pass through the origin at some values of δ, these trajectories form a group and their initial points may be called critical initial points. According to equations 7 and 8, the pair of equations defining the locus of the critical initial points are

$$U_{20} = -\frac{\delta T_m \tau_\phi}{J}\left(\frac{t}{\tau_\phi}-1+\epsilon^{-\frac{t}{\tau_\phi}}\right) \qquad (7)$$

$$U_{10} = -\frac{\delta T_m \tau_\phi^2}{J}\left(\frac{t^2}{2\tau_\phi^2}+1-\frac{t}{\tau_\phi}-\epsilon^{-\frac{t}{\tau_\phi}}\right) - U_{20}t \qquad (8)$$

Equation 8 may be rewritten as

$$U_{10} = \frac{\delta T_m \tau_\phi^2}{J}\left[\left(1+\frac{t}{\tau_\phi}\right)\left(\frac{t}{\tau_\phi}+\epsilon^{-\frac{t}{\tau_\phi}}-1\right)-\frac{t^2}{2\tau_\phi^2}\right] \qquad (9)$$

Starting from any point on the critical

curve defined by equations 7 and 9, a phase trajectory will pass through the origin.

Fig. 2 shows the critical curves on the phase plane. Curve 1 is obtained by setting δ=1, and curve 2 is obtained by setting δ=−1. The two curves join at the origin and divide the phase plane into two regions A and B. The optimum switching criterion is simply that δ=−1 in region A and δ=1 in region B.

Any large initial disturbance can be considered as a combined step input in both position and velocity, and corresponds to an initial point P on the phase plane. Suppose that this initial point falls in region A. The contactor operates to accelerate the motor (δ=−1), and the system follows the trajectory P-P' until it crosses the boundary at some

point on curve 1. There is no question that the phase trajectory will intersect with curve 1. As seen from equation 6, any phase trajectory will remain a phase trajectory on translating left or right. Starting at any point on curve 2, a phase trajectory would pass through the origin. Consequently, starting at any point to the right of curve 2, it will pass the U_1 axis at a point to the right of the origin, and subsequently intersect curve 1.

As soon as the phase trajectory crosses into region B, the contactor operates to decelerate the motor, and the system arrives at a steady condition following the phase trajectory P'-P'' of Fig. 2. Following approximately the same arguments as given in reference 1, it can be shown in general that along the two phase trajectories P-P' and P'-P'' the time required to reach the steady condition is a minimum.

Thus it follows that the optimum open-circuit switching criteria for a third-order system is a curve on the phase plane. In many ways it is similar to the second-order system described in reference 1. However, it differs from the latter in the following ways:

1. The critical boundary is not a trajectory.

2. For either value of δ, there are an infinite number of phase trajectories passing through a given point on the phase plane. The underlying reason for such is that
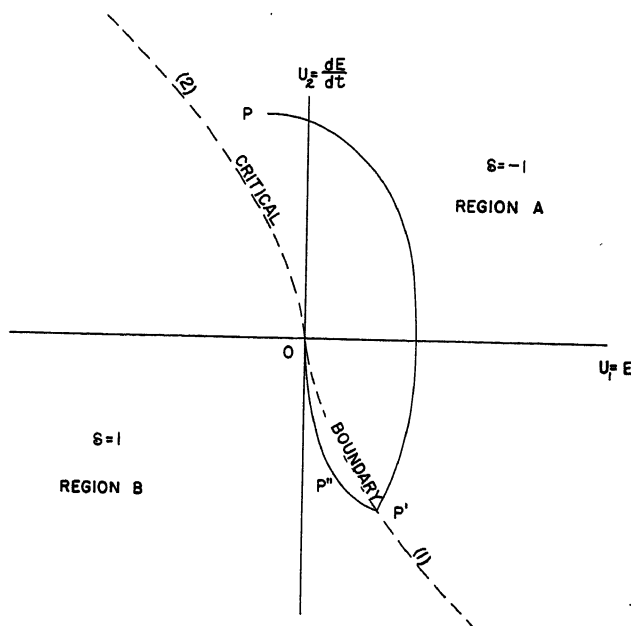
Fig. 2. Phase-plane illustration of function of computer with typical trajectories, curves P-P' and P'-P''-O
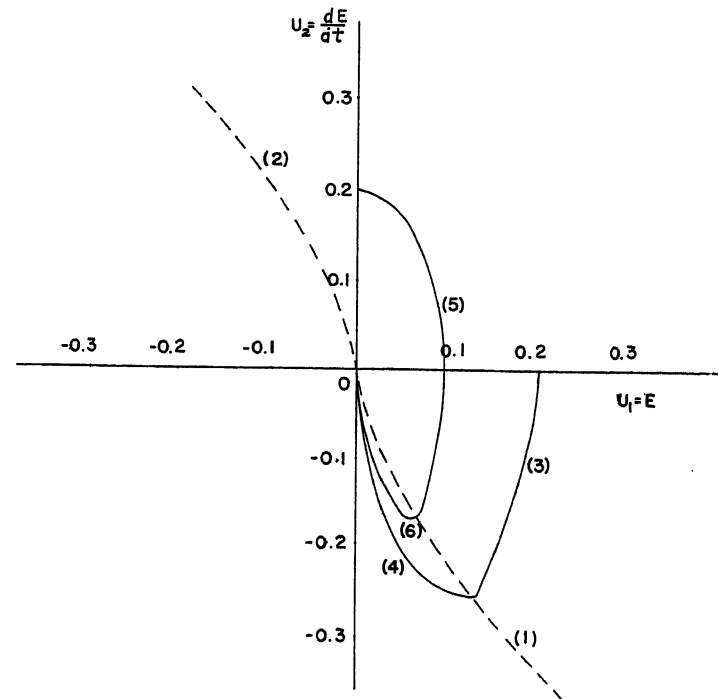
Fig. 3. Phase trajectories subsequent to step function position input, curves 3 and 4, and to step function velocity input, curves 5 and 6

U_1 is in units of $(T_m\tau_\phi^2)/J$

U_2 is in units of $(T_m\tau_\phi)/J$

Chang—Switching Criteria for Higher Order Contactor Servo

E axis graph (left). Curve with values 0.20, 0.16, 0.12, 0.08, 0.04 vertical; horizontal 0, 0.4, 0.8, 1.2, 1.6, 2.0, t/τ_ϕ.
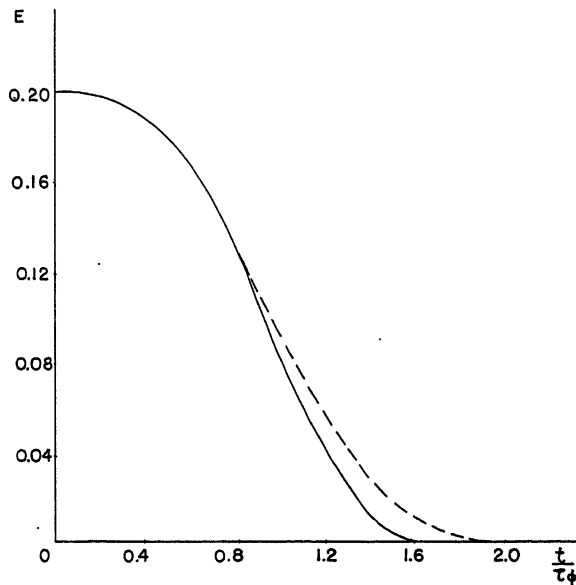
Fig. 4 (left). Error versus time with step function position input

Solid line—open-circuit switching
Dashed line—closed-circuit switching (reference 3)
Error E is in units of $(T_m\tau_\phi{}^2)/J$

Fig. 5 (right). Error velocity versus time with step function position input

Solid line—open-circuit switching
Dashed line—closed-circuit switching (reference 3)
dE/dt is in units of $(T_m\tau_\phi)/J$

$\frac{dE}{dt}$ axis graph (right). Vertical 0.16, 0.12, 0.08, 0.04, 0, −0.04, −0.08, −0.12, −0.16; horizontal 0.4, 0.8, 1.2, 1.6, 2.0, $\frac{t}{\tau_\phi}$.
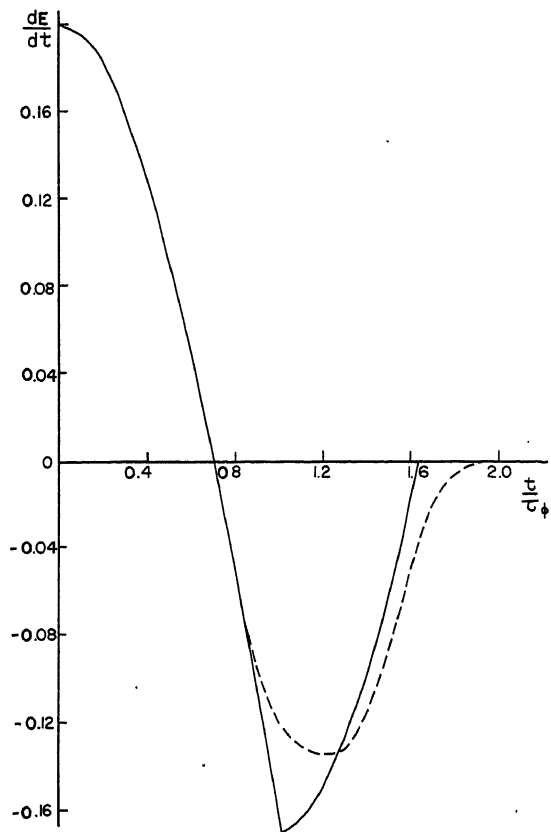
the dynamical state of a second-order system is completely defined by the position and velocity, while that of the third order system depends also on the torque which, in turn, is a function of the time after switching. Therefore, not only the point on the phase plane but also the time it takes to arrive at that point define a phase trajectory.

Curves 3 and 4 of Fig. 3 constitute the path on the phase plane which the system would follow with an initial step input of 0.2 unit in position. Similarly, curves 5 and 6 describe the system with an initial step input of 0.2 unit in velocity. The errors as a function of time are plotted in Figs. 4 and 5 for the respective cases. Compared with those for the same system described in reference 3, the present case takes about 20 per cent less time to reach a steady condition. Physically, the reason for such is quite clear. By quenching the field current with momentarily open-circuit switching, the time for reversing the field current is made shorter. However, more significant is the simplification in switching arrangements. The open-circuit switching requirement can be closely approximated by a circuit similar to that used in proportion plus derivative control, but with some nonlinear elements. Much more complicated arrangements would be necessary for optimum closed-circuit switching for the same system.

The General Case

While a contactor servo of fourth-order or higher is rarely met in practice, it may be of interest to study the general problem in its broader aspects. Suppose that in an N-order contactor servo the error E and its time derivatives up to the M-1 order are continuous ($M \le N-1$),

and its derivatives from the M-order to the $N-1$ order are zero at the moment of switching. The differential equation describing the system between switching intervals is

$$F\left(\delta, E, \frac{dE}{dt}, \dots \frac{d^n E}{dt^n}\right)=0 \qquad (10)$$

Generally, equation 10 can be solved if the initial values of E and its time derivatives up to the $N-1$ order are known. Since only M of these values are nonvanishing, the system can be described in an M-dimensional phase space. The co-ordinates of the phase space are defined as

$$U_r=\frac{d^{r-1}E}{(dt)^{r-1}} \quad r=1, 2, \dots M \qquad (11)$$

The equations describing a trajectory can be obtained by solving equation 10; subject to prescribed initial conditions they are

$$U_r = P_r(\delta, t, U_{1c}, U_{20}, \dots, U_{Mc})$$
$$r=1, 2, \dots M \quad (12)$$

The time variable t can be considered as a parameter. Each set of values of $U_{10}, \dots U_{M0}$ defines a distinct phase trajectory. For a given point U_{r1} in phase space, there are an infinite number of trajectories which pass through the given point at one time or the other. The initial points of these trajectories can be obtained by solving equation 12

for U_{r0} with $U_{r1}=U_r$. The solution can be written as

$$U_{r0}=Q_r(\delta, t, U_{11}, U_{21}, \dots, U_{M1})$$
$$r=1, 2, \dots M \quad (13)$$

The starting points from which a phase trajectory will eventually pass through the origin are given by

$$U_{r00}(\delta_0, t_0)=Q_r(\delta_0, t_0, 0, 0, 0, \dots 0), t_0>0$$
$$r=1, 2, \dots, M \quad (14)$$

As there is only one variable parameter, the M equations embodied in equation 14 specify a curve in the phase space. The two curves, with δ_0 equal to ± 1, join at the origin and form the final switching boundary.

The starting points from which a phase trajectory would pass through a point on the final switching boundary are given by

$$U_{r01}(\delta_1, t_0, t_1)=Q_r[-\delta_0, t_1, U_{100}(\delta_0 t_0),$$
$$\dots U_{M00}(\delta_0, t_0)] \quad r=1, 2, \dots M \quad (15)$$

As there are two variable parameter t_0 and t_1, equation 15 describes a surface starting from which a system can reach a steady condition with one switching operation. The surface forms that next-to-final switching boundary.

The foregoing method is repeated with the result that the points requiring two switching steps to reach balance form a 3-dimensional continuum, and the points needing three switching steps to reach balance form a 4-dimensional continuum, etc. Generally, the points taking $M-1$

switching steps would cover the entire M-dimensional phase space.

It has not been shown in this paper that solutions 12 and 13 exist and are unique. However, such follows from physical reasoning provided that the open-loop system is stable.

It may be concluded generally that, in a contactor servo of the N-order with the controlled variable continuous up to its $M-1$ derivative at the moment of switching, $M-1$ switching steps will be necessary to reach balance. The optimum switching criterion can be determined by successively applying equation 15. Compared to a high-order contactor servo with closed-circuit switching, its counterpart with open-circuit switching is faster in response and simpler in computer construction.

Appendix I. Derivation of Equation 1

The derivation of equation 1 is substantially the same as that given in reference 3. Referring to Fig. 1, the motor armature current I_g is assumed to be a constant, and the reversal of the torque is accomplished by reversing the field current i. Assuming that the field flux is proportional to i, the motor torque is

$$T = Ki \qquad (16)$$

Neglecting friction

$$T = J \frac{d^2\theta_0}{dt^2} \qquad (17)$$

Let the parameter δ be so defined that it is -1 if positive V is applied to the field circuit and vice versa, the field current i can be represented as

$$R_i + L \frac{di}{dt} = -\delta V \qquad (18)$$

Combining equations 16 and 18 one obtains

$$T + \tau_\phi \frac{dT}{dt} = -\delta T_m \qquad (19)$$

where

$$T_m = \frac{KV}{R}$$

$$\tau_\phi = \frac{L}{R}$$

Substituting equation 17 in equation 19

$$\tau_\phi \frac{d^3\theta_0}{dt^3} + \frac{d^2\theta_0}{dt^2} = -\delta \frac{T_m}{J} \qquad (20)$$

To evaluate the response of the system to an initial step input of position or velocity or any combination thereof, it may be assumed that no new disturbance is introduced before the first step change is fully compensated. Mathematically the assumption can be written as

$$\frac{d^n\theta_i}{dt^n} = 0 \quad n \geq 2 \qquad (21)$$

Equation 1 is obtained by combining equations 20 and 21.

References

1. A Phase-Plane Approach to the Compensation of Saturating Servomechanisms, Arthur M. Hopkin. *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 631-39.

2. Nonlinear Techniques for Improving Servo Performance, D. McDonald. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 6, 1950, pp. 400-21.

3. An Investigation of the Switching Criteria for Higher Order Contactor Servomechanisms, Irving Bogner, Louis F. Kazda. *AIEE Transactions*, vol. 73, pt. II, July 1954, pp. 118-27.

Short-Circuit Currents in Low-Voltage Systems

JOINT SECTIONS COMMITTEE ON AIR CIRCUIT BREAKERS OF NEMA

ONE of the aims of the Joint Sections Committee on Air Circuit Breakers of the National Electrical Manufacturers Association (NEMA) is to propose standards of satisfactory performance for low-voltage (lv) air circuit breakers under short-circuit conditions. To determine the requirements for satisfactory circuit-breaker short-circuit performance, it is necessary to obtain information about the various short-circuit conditions that exist in electric power systems where the circuit breakers may be installed. Sponsored by the NEMA Joint Sections

Committee, studies have been made to permit the maximum probable short-circuit duties to be determined at nearly any point in a wide range of low-voltage electric power systems. This paper gives the basic assumptions and procedures used in the calculations and presents the results as a series of curves that can be used to determine short-circuit duties in lv systems.

The study covers the most common lv systems, which are radial, 3-phase systems operating at 60 cycles per second and supplied through transformers. Network systems and systems with local generation are not included in the general evaluation because they are usually special cases and require individual consideration.

Fig. 1 shows the radial circuit employed in investigating lv short-circuit duties. Solutions for the symmetrical current at the point of short circuit are based on the methods described in AIEE Standard C37.5[1] and in the supplement to AIEE

Standard C37.13.[2] To be conservative, maximum probable short-circuit current values are calculated by selecting circuit impedances that are minimum probable values, as described later, and all calculations are based on bolted faults.

Physically, as Fig. 1 indicates, the short-circuit duty depends on the following items: available primary short-circuit duty, transformer size and impedance, system voltage, sizes and effective impedances of motors contributing to the short-circuit current, and the size, length, and cross-section geometry of the feeder conductors connecting the lv source bus to the point at which the short-circuit duty is being investigated.

Eight curve sheets, Figs. 2 through 9, present the results of the calculations, one for each of the following transformer kilovolt-ampere (kva) sizes: 150, 225, 300, 500, 750, 1,000, 1,500 and 2,000. The independent variable plotted on the horizontal scale of each graph is the

Fig. 1. Circuit investigated and results obtained
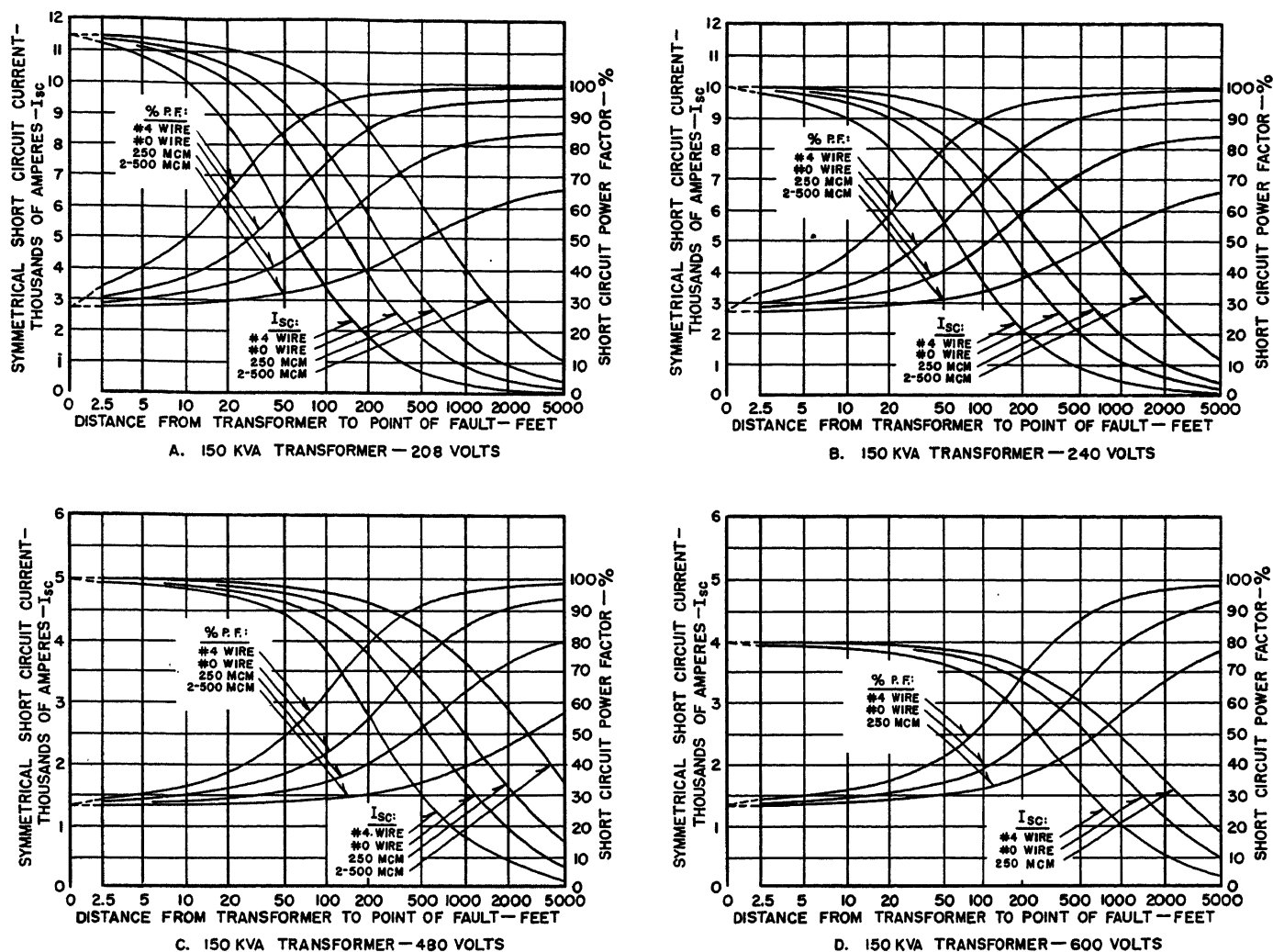
Fig. 2. Symmetrical short-circuit current and power factor versus distance from a 150-kva liquid-filled power transformer: X/R=3.24; R=1.23 per cent; X=4.0 per cent; and Z=4.19 per cent

ength in circuit feet of the feeder conductors connecting the lv source bus and the point of short circuit. Five different sizes and feeder conductor cross-section arrangements have been selected and investigated.

The five conductor arrangements, originally chosen to be at least 115 per cent of the size normally required for the standard frame sizes of lv air circuit breakers are given in Table I. The meaning of these designations is explained by example: Four 750,000 circular mils (CM) indicates that the feeder conductors consist of four 3-conductor 750,000-CM cables operating in parallel with one conductor of each cable carrying one-fourth of the total current of each phase. Feeder arrangements that have essentially the same short-circuit duties for equal length of conductors can be substituted, as shown in Table II, for feeder arrangements in the fourth and fifth lines of Table I respectively.

Approximate short-circuit current and power-factor values for other feeder

conductor sizes and arrangements can be obtained by interpolating between the lines on any curve sheet, using impedance values as a rough basis for the interpolation. The results of the calculations displayed graphically permit determining at any point in a given lv system two factors that will define the short-circuit duty at that point. One is the magnitude of the a-c rms symmetrical component of the short-circuit current and the other is the power factor of the short-circuit current. The total asymmetrical short-circuit current may be determined from the plotted symmetrical short-circuit current by using a multiplier that is a function of short-circuit power factor, as shown in Fig. 10.

Basic Impedance Data and System Conditions Used for Calculations

As mentioned previously, the equivalent circuit investigated consists of impedances that are considered minimum values. These are described here.

PRIMARY SOURCE CIRCUIT

It is assumed that a 500-megavolt-ampere (mva) short-circuit duty is available at the primary of the transformer, and that the source circuit reactance-resistance (X/R) ratio is 25. This X/R ratio corresponds roughly to the standard multiplier of 1.6 used to obtain maximum phase rms total currents in primary circuits from calculated symmetrical current values.

It is believed that very few primary circuits have an available short-circuit duty that is greater than 500 mva with an X/R of 25. Even if the available pri-

Table I. Conductor Arrangements

Rating of Circuit, Amperes	Conductor Size	Type of Conductor Insulation
50	no. 4	R
100	no. 0	R
225	250,000 CM	RH
600	two 500,000 CM	RH
1,600	four 750,000 CM	RH
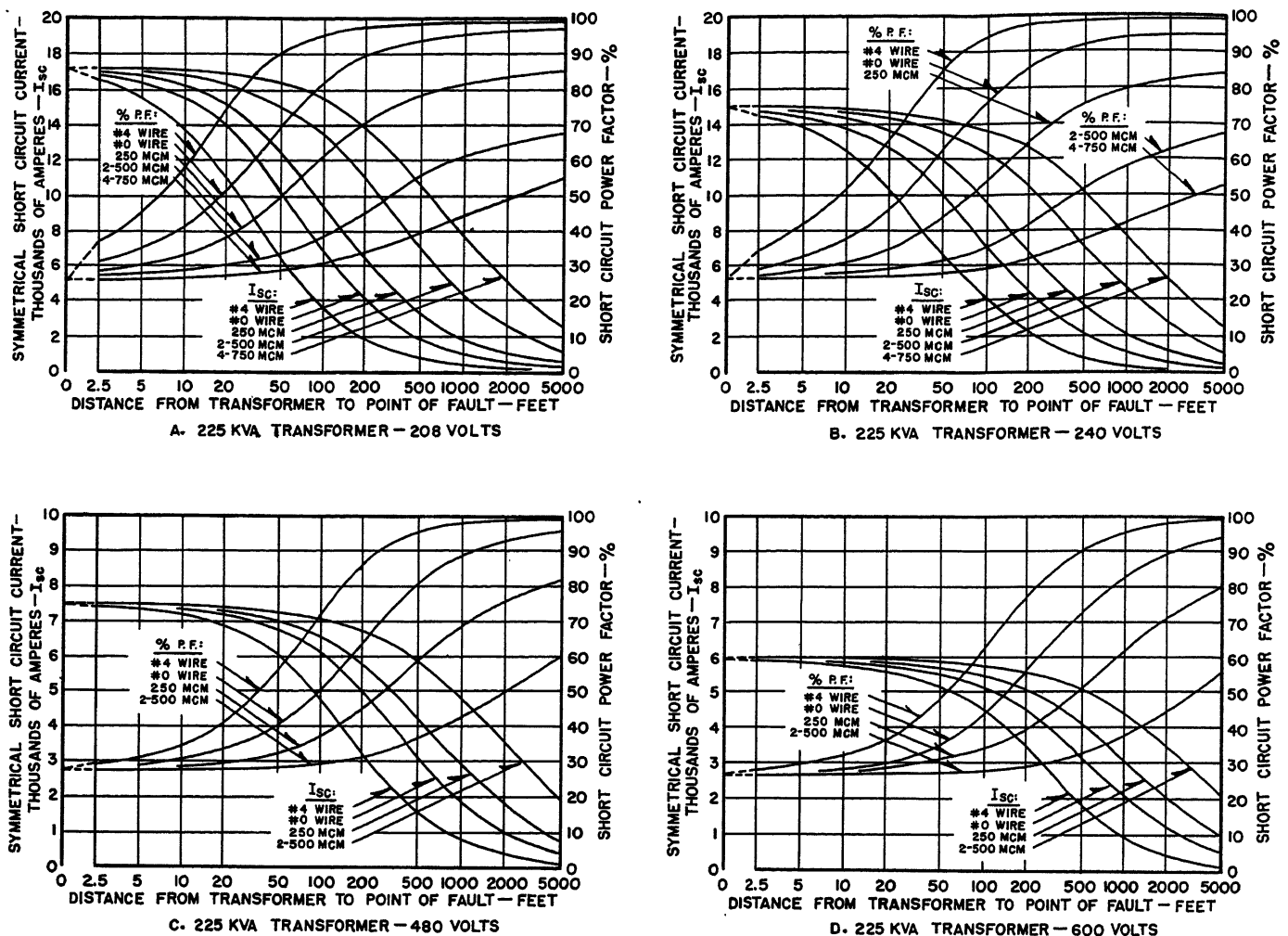
Fig. 3. Symmetrical short-circuit current and power factor versus distance from a 225-kva liquid-filled power transformer: X/R = 3.35; R = 1.19 per cent; X = 4.0 per cent; and Z = 4.17 per cent

mary duty exceeds this, it is probable that the plotted results would still prove to be conservative because there are other impedances in the circuit, also selected at minimum values, and the probabilities are very slight that any one system will have all impedances at the minimum possible values concurrently.

TRANSFORMER IMPEDANCES

To determine the transformer characteristics, nominal standard transformer impedances of $4\frac{1}{2}$ per cent for transformers having ratings up to and including 500 kva, and $5\frac{1}{2}$ per cent for transformers having ratings above 500 kva are assumed. The allowable standard tolerance of $-7\frac{1}{2}$ per cent is then applied to select a minimum preliminary value for the impedance Z of each transformer size. From data on liquid-filled, self-cooled transformers supplied by numerous transformer manufacturers, X/R ratios are selected for each transformer size. From the minimum preliminary impedances and the X/R ratios, values of X are calculated and then slightly

rounded off. From the rounded off values of X and the X/R values, new values of R and Z are calculated for each transformer. Table III shows a tabulation of the impedance information for sizes of transformers investigated in the study.

FEEDER CONDUCTOR DATA

Resistance and reactance are computed for 3-phase cables, having only the minimum allowable thickness of insulation around each conductor, installed in a magnetic duct or having a magnetic binder such as steel-interlocked armor. The values as calculated are almost exactly equivalent to those for 3-phase cable with minimum insulation plus a braid around each conductor installed in nonmagnetic duct or in free air.

The impedance as computed is almost the minimum possible impedance for the conductor sizes considered and is, therefore, conservative. Reactances would be higher with insulation thicker than the standard minimum or with three single conductors instead of a 3-

conductor cable pulled into a conduit because, in both cases, the average inter-conductor spacing is greater. Therefore, the curves may be used for circuit conductors having the same wire size, but a different arrangement. Table IV shows the impedance information for the cable circuits investigated and, in addition, the standard minimum insulation thickness.

MOTOR IMPEDANCES

It is assumed that the short-circuit current includes contributions from a group

Table II. Substitute Feeder Arrangements for Which Short-Circuit Current and Power Factor Values May be Read from the Curves

Conductor Arrangement per Phase for Which Curve Is Drawn	Substitute Conductor Arrangement per Phase for Which Curve Values May Be Used
Two 500,000 CM	three no. 4/0 American Wire Gauge
	four no. 2/0 American Wire Gauge
Four 750,000 CM	three 2,000,000 CM
	five 400,000 CM
	six 300,000 CM
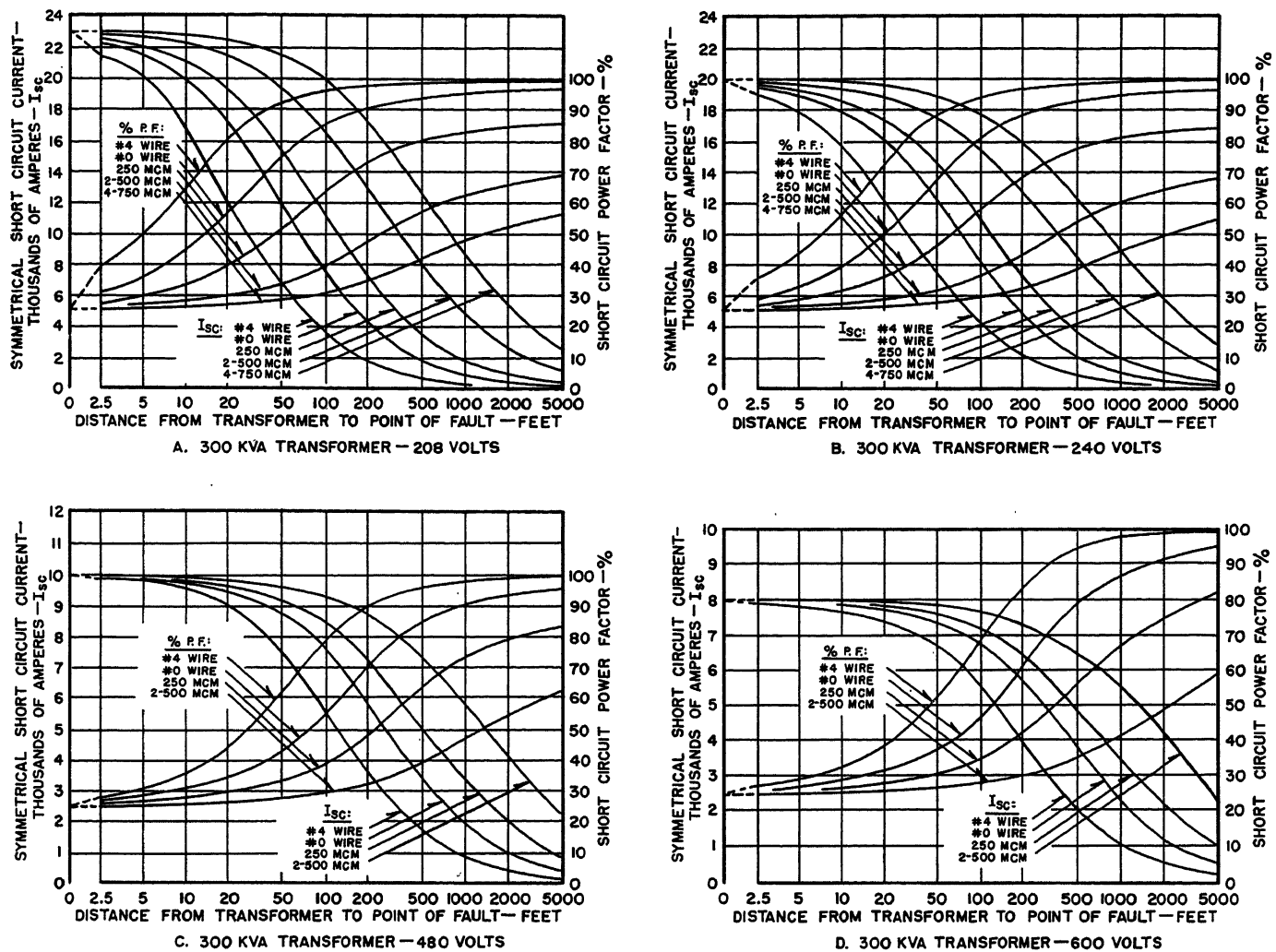
Fig. 4. Symmetrical short-circuit current and power factor versus distance from a 300-kva liquid-filled power transformer: X/R = 3.50, R = 1.14 per cent, X = 4.0 per cent, and Z = 4.16 per cent

of motors connected to the transformer secondary bus, and that the total kva of the group of motors is equal to the transformer kva. The group of motors is assumed to have an equivalent reactance of 25 per cent and an X/R of 6. These values are approximately correct for terminal short circuits of individual 30-horsepower induction motors, are conservative for groups of smaller horsepower motors, and satisfactorily represent larger motors with short lengths of cable circuit connecting the motors to the transformer bus.

Using the Curves to Find Short-Circuit Duties

The curves, Figs. 2 through 9, are used directly to find the short-circuit current duties and power factor. To find a short-circuit current at the end of the given length of a given size feeder originating at the lv bus of a given kva substation; on the appropriate graph and on the short-circuit current curve

marked with the feeder size, find a point corresponding to the length of the feeder as read on the horizontal scale, then find the short-circuit current corresponding to that point on the vertical current scale. Power factors are found in a similar manner.

Where the circuit from the substation lv bus includes two sections of different wire size, an estimate of the duty can be obtained considering only one wire size. First, find the short-circuit current corresponding to the length and size of

Table III. Assumed Transformer Characteristics for all Applicable Voltages

Transformer, Kva	X/R	R, Per Cent	X, Per Cent	Z, Per Cent
150	3.24	1.23	4.0	4.19
225	3.35	1.19	4.0	4.17
300	3.50	1.14	4.0	4.16
500	3.84	1.04	4.0	4.12
750	5.45	0.94	5.1	5.19
1,000	5.70	0.89	5.1	5.19
1,500	6.15	0.83	5.1	5.18
2,000	6.63	0.77	5.1	5.17

the section having the smaller wire. Then find the short-circuit current corresponding to the wire size of the other section, but using a length equal to the total distance, i.e., the sum of the two section lengths. Since the actual duty is less than either of the two short-circuit currents thus determined, use the smaller current value.

ALTERNATIVE METHOD TO FIND SHORT-CIRCUIT DUTIES

As an alternative, the short-circuit current and power factor for any length of feeder at any system voltage may be computed from the curves for one voltage by using the following procedure. Take the curves for the 480-volt system as a base (lower left-hand curves on Figs. 2 through 9).

1. Short-circuit current I_{sc} at any system voltage V_s at feeder length L may be found on the 480-volt curves as follows:

a. Choose the 480-volt curve of the proper kva transformer under study.

b. Select a distance along the abscissa

Fig. 5. Symmetrical short-circuit current and power factor versus distance from a 500-kva liquid-filled power transformer: X/R=3.84; R=1.04 per cent; X=4.0 per cent; and Z=4.12 per cent

which will be $L_{480\ volts}=L$ (at system voltage)$\times(480/V_s)^2$. The factor $(480/V_s)^2$ for common service voltages is given in Table V.

c-1. Read the short-circuit current on the 480-volt curve for the proper feeder at the distance found in step 2.

c-2. The short-circuit current at the desired system voltage will then be I_{sc} (at system voltage)$=I_{sc}$ (from c-1)$\times(480/V_s)$. The factor $(480/V_s)$ for common service voltages is given in Table VI.

2. The short-circuit power factor can be read directly from the 480-volt curve corresponding to the proper feeder size at a distance as found in step 1-b.

Sample calculations for a bus short-circuit duty and for short-circuit current and power factor are given in Appendixes I and II respectively.

Sample calculations and examples illustrating the use of the curves will be found in Appendix III. The examples in Appendix III show how the full set of curves is used directly, the alternative method of computation for any system voltage from the set of 480-volt curves,

and the short-circuit current calculation in a feeder system involving two sizes of conductors.

ASYMMETRICAL CURRENT VALUES

The currents read from the graph and computed in accordance with the methods and equations used in the previous paragraph are all symmetrical currents. To find the asymmetrical values corresponding to the symmetrical values, the

curve shown as Fig. 10 is used. As previously explained, the degree of asymmetry can be related to the power factor of the short circuit.

In present-day usage, the asymmetrical current value compared with lv circuit-breaker ratings is computed by averaging the three total rms asymmetrical currents at 1/2 cycle of a 3-phase bolted short circuit. This average short-circuit asymmetrical current can be found by multi-

Table IV. Cable Data

Rating of Circuit, Amperes	Wire Size	Characteristics per Conductor		
		Resistance per 100 Feet*	Reactance per 100 Feet	Thickness of Insulation, Inches
50	No. 4	0.0312	0.00362	4/64
100	No. 0	0.0125	0.00340	5/64
225	250,000 CM	0.00547	0.00322	6/64
600	Two 500,000 CM	0.00292†	0.00295†	6/64
1,600	Four 750,000 CM	0.00208†	0.00284†	7/64

* Resistance corresponds to 75 degrees centigrade corrected for 60 cycles a-c, 600-volt cable.

† Values tabulated are for single conductors. Calculations for 600- and 1,600-ampere circuits (2 and 4 conductors respectively) are obtained by paralleling these values without consideration of more than one cable in single conduit.
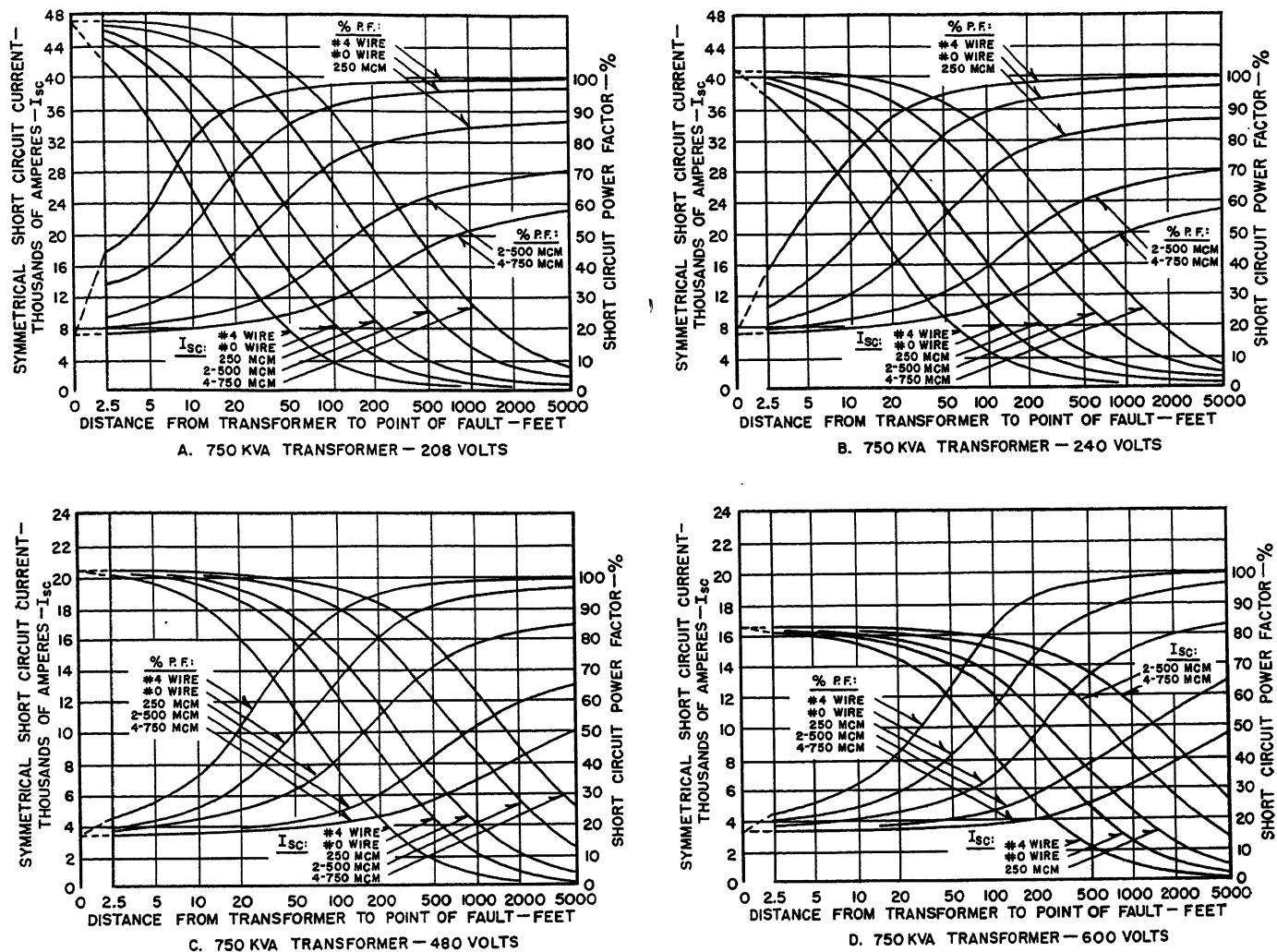
Fig. 6. Symmetrical short-circuit current and power factor versus distance from a 750-kva liquid-filled power transformer: X/R=5.45; R=0.94 per cent; X=5.1 per cent; and Z=5.19 per cent

Table V. Distance Factor $(480/V_s)^2$ for Common Service Voltages

Voltage	$\left(\dfrac{480}{V_s}\right)^2$
208	5.34
240	4.00
480	1.00
550	0.76
600	0.64

Table VI. Short-Circuit Current Factor $(480/V_s)$ for Common Service Voltages

Voltage	$\left(\dfrac{480}{V_s}\right)$
208	2.31
240	2.00
480	1.00
550	0.87
600	0.80

plying the symmetrical current determined from the graphs by the multiplier M_A, shown in Fig. 10, corresponding to the short-circuit power factor. If it is desired to find the maximum possible

rms asymmetrical current among the three that are averaged for the value found previously, then use the multiplier M_M found in Fig. 10 corresponding to the power factor.

Appendix I. Sample Calculation for a Bus Short-Circuit Duty

A sample per-unit calculation for the bus short-circuit current and for the equivalent impedance to the bus is given for the 750-kva transformer size; note that the bases are 1 mva=1,000 kva, 480 volts, 1,203 amperes.

Source equivalent impedance (500 mva, $X/R=25$), $Z_S=0.00008+j0.0020$ (1)

Transformer impedance

$$=\left(\frac{0.94\text{ per cent}}{100}+j\frac{5.1\text{ per cent}}{100}\right)\frac{1,000}{750},$$

$$Z_T=0.01253+j0.0680 \quad (2)$$

Sum of source+transformer impedances, $Z_{T+S}=0.01261+j0.0700$ (3)

Motor impedance (750 kva, $X=25$ per cent,

$$X/R=6)=\frac{1,000}{750}(1/6+j1)0.25,$$

$$=Z_M=0.0555+j0.333 \quad (4)$$

Source and transformer admittance, $\dfrac{1}{Z_{T+S}}$

=per-unit current contribution,

$$(T+S)=Y_{T+S}=2.493-j13.81 \quad (5)$$

Motor admittance, $\dfrac{1}{Z_M}$=per-unit current

contribution, $(M)=Y_M=0.486-j2.92$ (6)

Total per-unit symmetrical current, bus short circuit $=Y_{T+S}+Y_M=2.979-j16.73$ (7)

Reciprocal of equation 7 $Z_{(bus)}=0.01029+j0.0579$ (8)

Polar form of equation 7=per-unit $I_{sc(bus)}$
$$\angle\theta°=17.00\ \angle-79.91° \quad (9)$$

Short-circuit power factor(per cent)
$$=\cos\theta°\times100=17.5\text{ per cent} \quad (10)$$

Symmetrical $I_{sc\ (bus)}$ at 480 volts
$$=1,203(i)=20,450\text{ amperes} \quad (11)$$

A. 1000 KVA TRANSFORMER – 208 VOLTS

B. 1000 KVA TRANSFORMER – 240 VOLTS

C. 1000 KVA TRANSFORMER – 480 VOLTS

D. 1000 KVA TRANSFORMER – 600 VOLTS
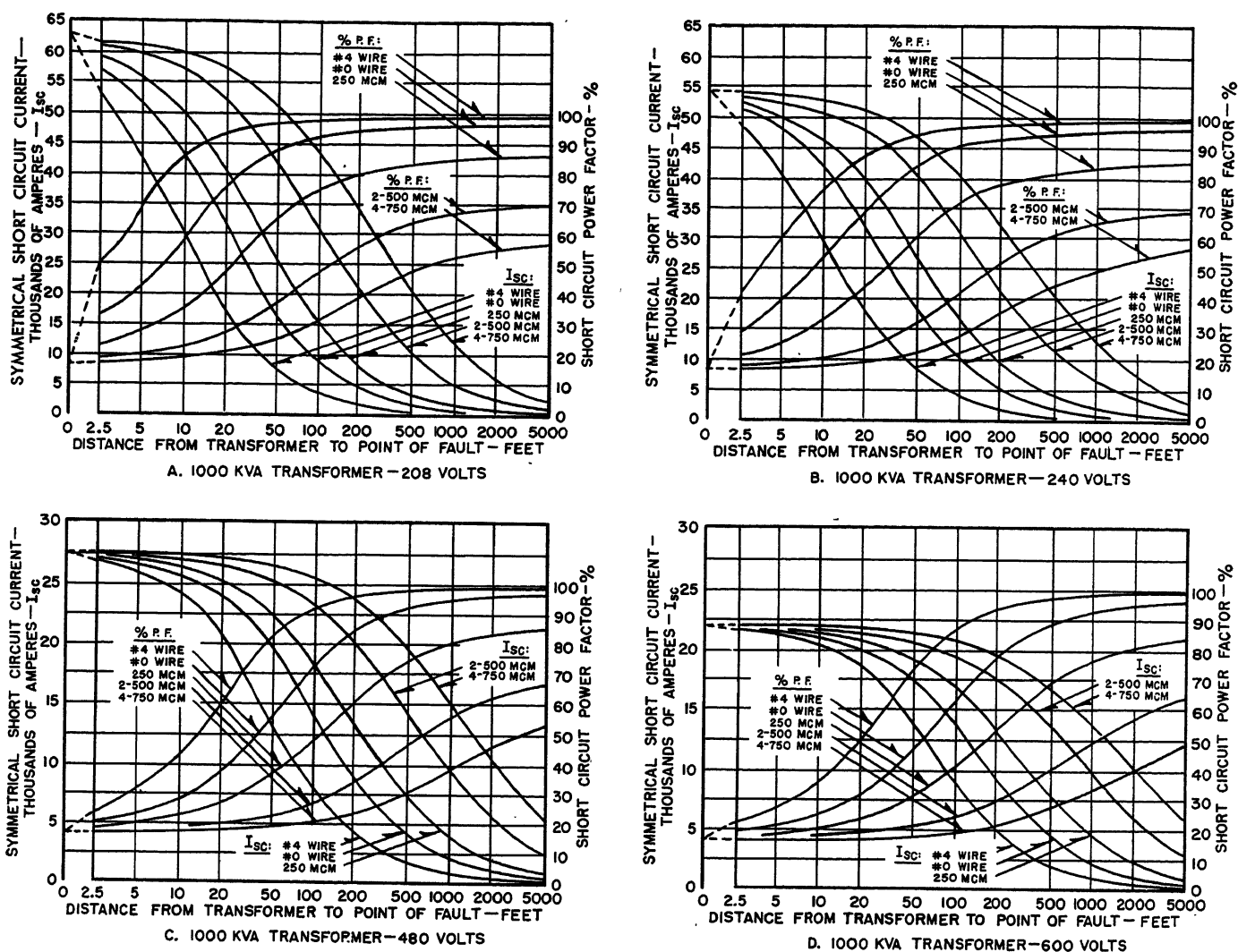
Fig. 7. Symmetrical short circuit current and power factor versus distance from a 1,000-kva liquid-filled power transformer: $X/R = 5.70$, $R = 0.89$ per cent, $X = 5.1$ per cent, and $Z = 5.19$ per cent

Symmetrical I_{sc} (bus) at 208/120 volts
$$= 2.31(k) = 47,200 \text{ amperes} \quad (12)$$

Symmetrical I_{sc} (bus) at 240 volts $= 2.00(k)$
$$= 40,900 \text{ amperes} \quad (13)$$

Symmetrical I_{sc} (bus) at 600 volts $= 0.80(k)$
$$= 16,360 \text{ amperes} \quad (14)$$

See Table VI for multiplier used in equations 12, 13, and 14.

Appendix II. Sample Calculation of Short-Circuit Current and Power Factor

Here is a sample calculation of the short-circuit current and power factor on a 225-ampere feeder, 320 feet from a 750-kva 480-volt transformer substation.

System impedance to low voltage transformer bus (from Table VII)
$$= Z_{\text{(bus)}} = 0.01029 + j0.0579 \quad (15)$$

Impedance of 320-foot feeder ($3.2 \times$ value from Table VIII) $= Z_{\text{(feeder)}}$
$$= 0.0760 + j0.0447 \quad (16)$$

Total impedance to fault, $Z_{\text{(bus)}} + Z_{\text{(feeder)}}$
$$= Z_T = 0.0863 + j0.1026 \quad (17)$$

Per-unit short-circuit current $= 1/Z_T$
$$= I_{sc} \angle \theta^\circ = 7.46 \angle -49.93^\circ \quad (18)$$

Short-circuit power factor $\cos \theta^\circ \times 100$
$$= 64.4 \text{ per cent} \quad (19)$$

Symmetrical short-circuit current (in amperes) $1,203(d) = 8,970 \text{ amperes}$
$$(20)$$

Appendix III. Sample Calculations and Examples Illustrating the Use of Curves

Example 1

Determine the short-circuit current available and the power factor on a system voltage of 208 volts, connected to a liquid-filled transformer of 1,000 kva with 100 feet of 250,000-CM feeder conductor to the

Table VII. Bus Equivalent Impedances and Bus Short-Circuit Current Duties

Transformer, Kva	Equivalent Per-Unit Impedance to Bus, Including Motors, Z Bus (1,000-Kva Base)	Per-Unit Symmetrical Short-Circuit Current (1,000-Kva Base)	Short-Circuit Power Factor (All Voltages), Per Cent	Symmetrical Short-Circuit Current, Amperes with System Voltages of			
				208 Volts	240 Volts	480 Volts	600 Volts
150	0.066 + j0.2315	4.15	27.4	11,500	9,980	4,990	3,990
225	0.0428 + j0.1551	6.21	26.6	17,220	14,940	7,470	5,980
300	0.0308 + j0.1164	8.30	25.6	23,000	19,970	9,985	7,990
500	0.01712 + j0.0706	13.76	23.6	38,200	33,100	16,550	13,230
750	0.01029 + j0.0579	17.00	17.5	47,200	40,900	20,450	16,360
1,000	0.00736 + j0.0436	22.60	16.62	62,700	54,400	27,200	21,750
1,500	0.00467 + j0.0296	33.33	15.57	92,400	80,100	40,050	32,050
2,000	0.00332 + j0.0226	43.90	14.54	121,800	105,600	52,800	42,200

A. 1500 KVA TRANSFORMER—208 VOLTS

B. 1500 KVA TRANSFORMER—240 VOLTS

C. 1500 KVA TRANSFORMER—480 VOLTS

D. 1500 KVA TRANSFORMER—600 VOLTS
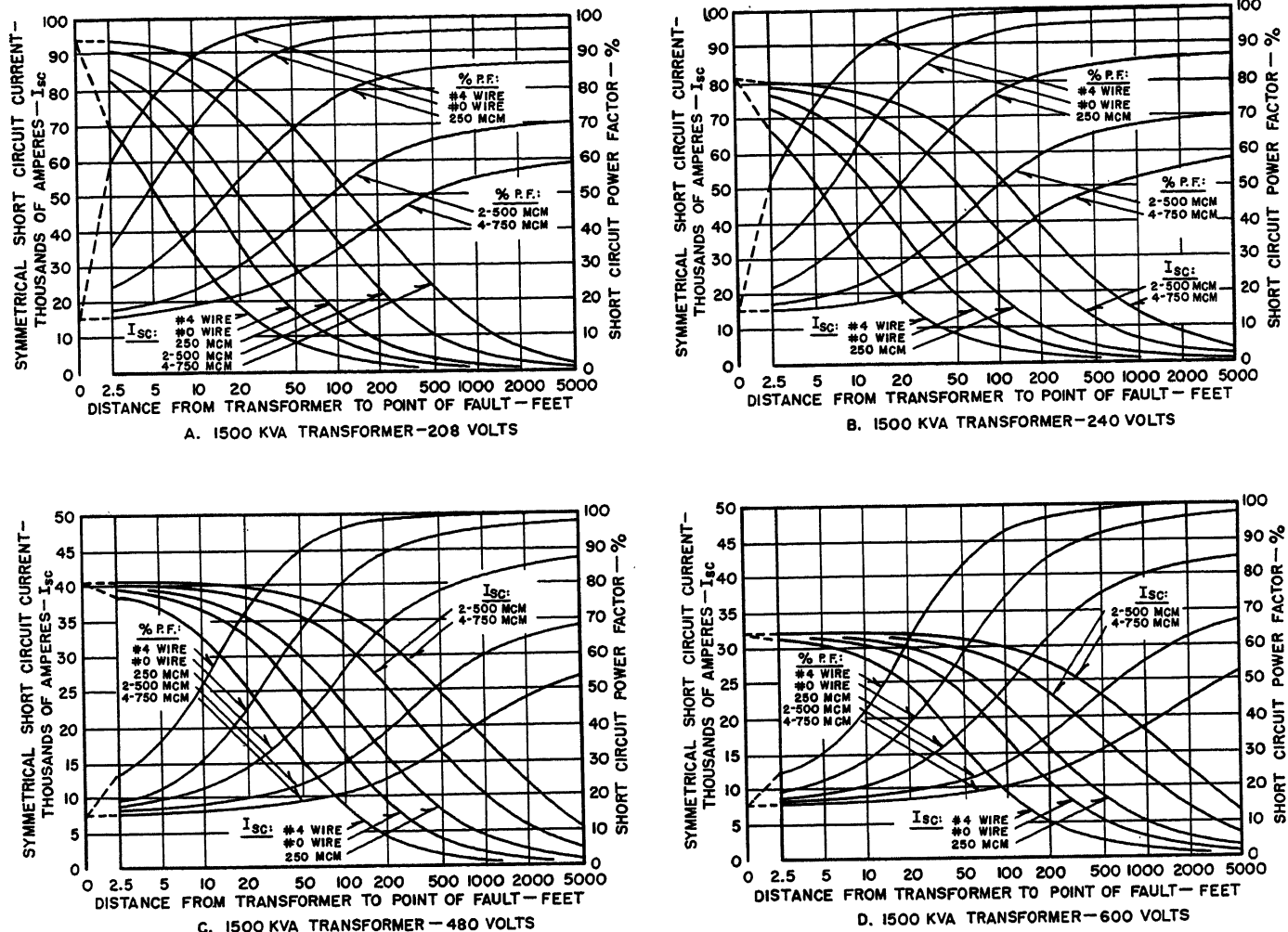
Fig. 8.   Symmetrical short-circuit current and power factor versus distance from a 1,500-kva liquid-filled power transformer: X/R=6.15;   R=0.83 per cent;   X=5.1 per cent;   and Z=5.18 per cent

point of fault. In Fig. 7(A) enter at 100 feet and read 15,400 symmetrical amperes at 75-per-cent power factor.

Example 2

Determine the short-circuit current available and the power factor on a system voltage of 208 volts connected to a liquid-filled transformer of 1,000 kva with 100 feet of 250,000-CM feeder conductor to the point of fault, from the alternative method, using the 480-volt curves.

1. Refer to Fig. 7(C).
2. $L_{(480\ volts)} = 100 \times 5.32 = 532$ feet. The 5.32 factor is taken from Table V.
3. Enter curve at 532 feet and read 6,700 amperes on the 250,000-CM curve.
4. The short-circuit current at 208 volts will then be I_{sc} (at 208 volts)= $6,700 \times 2.31 = 15,400$ amperes (symmetrical). The 2.31 factor is taken from Table VI.
5. The short-circuit power factor can be read directly from the same curve. Enter at 532 feet and read 75-per-cent power factor for the 250,000-CM conductor.

Example 3

Determine the short-circuit current available on a system voltage of 480 volts con-

nected to a liquid-filled transformer of 500 kva with two sections of different size conductor to the point of fault, the first section nearest the transformer consisting of 10 feet of two 500,000-CM conductors, and the other section beyond consisting of 50 feet of no. 0 wire. An estimate of the available short-circuit current can be obtained by considering only one conductor size.

1. From Fig. 5(C) read the current at 50 feet (length of the smaller conductor only) for the no. 0 wire. This is 13,500 symmetrical amperes.
2. Next, from Fig. 5(C) read the current at 60 feet (combined length of the two con-

ductors) for two 500,000-CM conductors. This is 15,500 symmetrical amperes.
3. Since the actual duty will be less than either of these figures, use the smaller one, or 13,500 symmetrical amperes.

References

1. METHODS FOR DETERMINING THE RMS VALUE OF A SINUSOIDAL CURRENT WAVE AND A NORMAL FREQUENCY RECOVERY VOLTAGE AND FOR SIMPLIFIED CALCULATION OF FAULT CURRENT. *ASA Standard C37.5*, American Standards Association, New York, N. Y., Dec. 1953.

2. LOW-VOLTAGE AIR CIRCUIT BREAKERS (including Application Guide). *ASA Standard C37.13, Ibid.*, Aug. 1954.

Table VIII.   Per-Unit Impedances of 100-Foot Feeder Circuits;   Base Kva = 1,000, Base Line-to-Line Voltage = 480, Base Line Current = 1,203

Rating of Feeder Circuit, Amperes	Conductors per Phase	Ohms per 100 Feet	Per-Unit Impedance per 100 Feet
50	no. 4 American Wire Gauge	$0.0312 + j0.00362$	$0.1353 + j0.01571$
100	no. 1/0 American Wire Gauge	$0.0125 + j0.00340$	$0.0542 + j0.01476$
225	250,000 CM	$0.00547 + j0.00322$	$0.02375 + j0.01397$
600	two 500,000 CM	$1/2(0.00292 + j0.00295)$	$0.00633 + j0.00640$
1,600	four 750,000 CM	$1/4(0.00208 + j0.00284)$	$0.002255 + j0.00308$

A. 2000 KVA TRANSFORMER—208 VOLTS

B. 2000 KVA TRANSFORMER—240 VOLTS

C. 2000 KVA TRANSFORMER—480 VOLTS

D. 2000 KVA TRANSFORMER—600 VOLTS
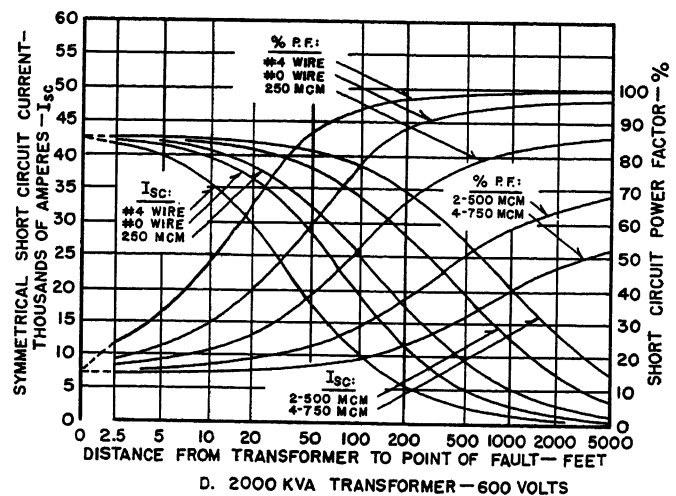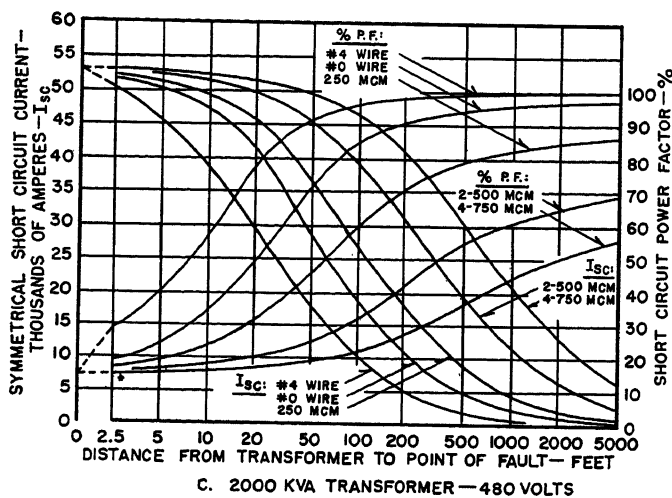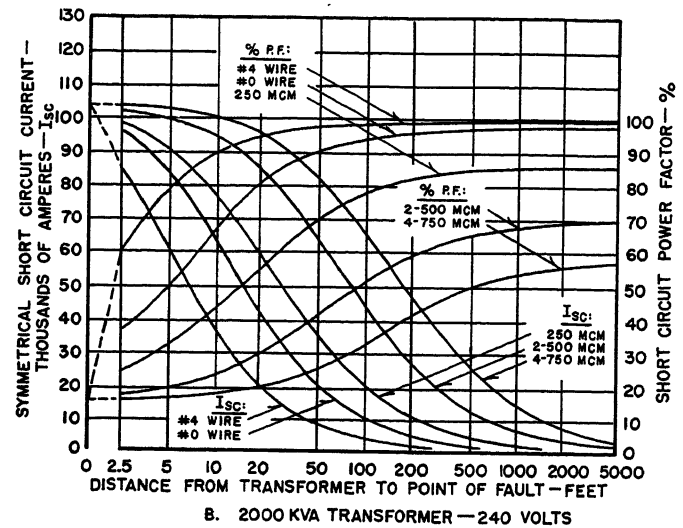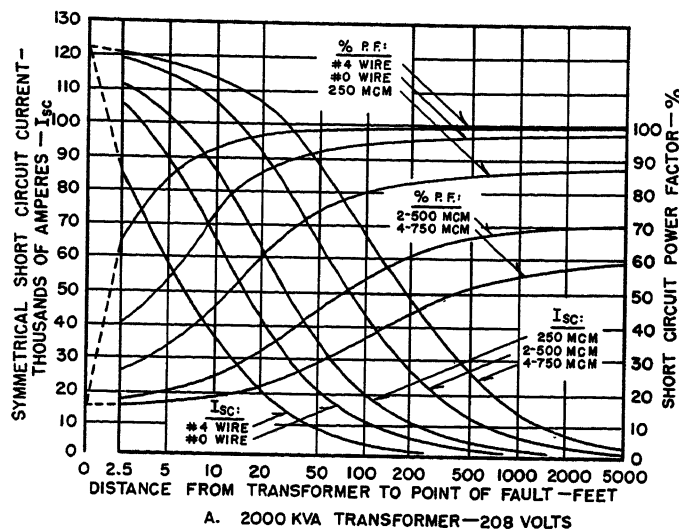
Fig. 9 (above). Symmetrical short-circuit current and power factor versus distance from a 2,000-kva liquid-filled power transformer: X/R=6.63; R=0.77 per cent; X=5.1 per cent; and Z=5.17 per cent

$$M_M = \frac{\text{TOTAL ASYMMETRICAL RMS AMPERES IN MAX. PHASE}}{\text{SYMMETRICAL RMS AMPERES}}$$

$$M_A = \frac{\text{AVE. 3 PHASE ASYMMETRICAL RMS AMPERES}}{\text{SYMMETRICAL RMS AMPERES}}$$
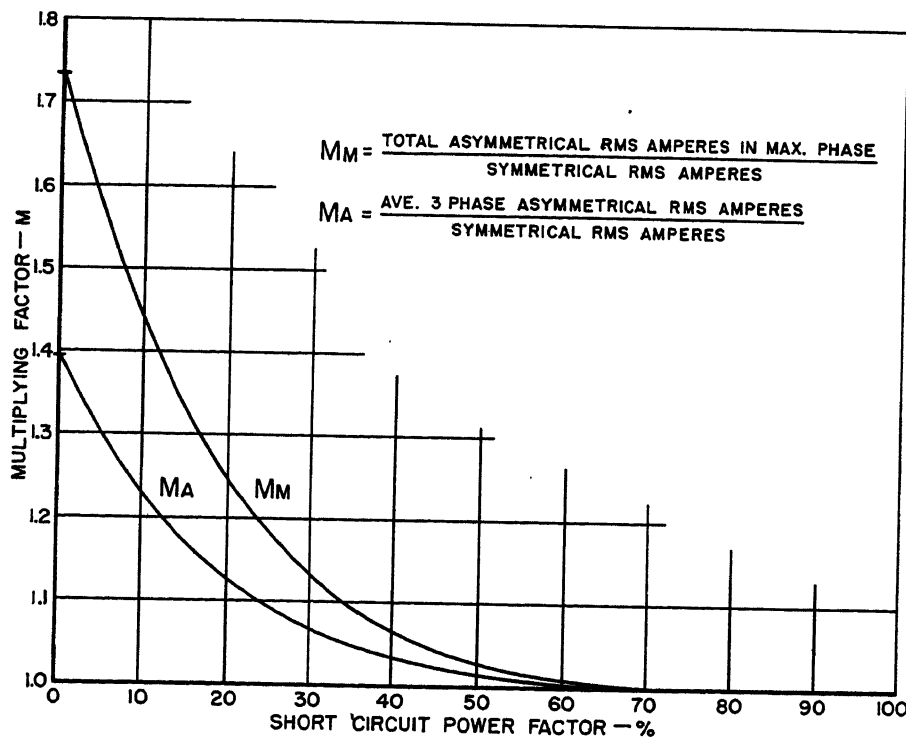
Fig. 10. Multiplying factors to obtain short-circuit asymmetrical current from symmetrical values, at an instant 1/2 cycle (60-cycle basis) after initiation of a fault

Discussion

Dan Clark (Bureau of Reclamation, Denver, Colo.): The data presented in this paper are complete, interesting, and believed to include all the material necessary to formulate a very simple guide for determining the short circuits available at any point in an lv system. A guide of this type would be very useful in determining the proper-size circuit breakers required for a system.

The basic impedance data and system conditions used for the calculations are believed to be more applicable to industrial plant power systems than to station service systems for hydroelectric power plants. The assumed total kva connected motor load which is equal to the transformer kva capacity appears to be too high for power-plant systems. Therefore, the assumed motor contribution to the fault current is too high. The transformer impedances used in the calculations are minimum probable values. The use of these conservative data may possibly make the cost of the switchgear for a power plant more expensive

than necessary, provided the circuit breakers were larger than actually required. For some power-plant station service systems, depending on the transformer capacity, it is more economical to purchase additional impedance in the transformer than to purchase the larger size switchgear. In addition to a considerable savings in the cost of the switchgear, the savings in space requirements of the gear must also be evaluated.

It would be interesting to see a few comparative tables and curves based on more nearly average conditions for power-plant station service systems using unit substations with radial feeders.

Arnold R. Kelly (Esso Research and Engineering Company, Linden, N. J.): The authors allow for motor contribution to fault by assuming an equivalent motor, equal in size to the transformer kva, and with a reactance of 25 per cent, an X/R of six, and no subtransient decay of a-c contribution. Equivalent impedance would be 25.3 per cent.

Traditional allowance for induction motor contribution during the first half-cycle of fault, when the size and number of motors justifies, consists of assuming an equivalent motor with admittance equal to the sum of the individual motor locked-rotor admittances and no subtransient a-c decay. Although connecting cable admittance may be included and the sum for all motors added as a phasor quantity, the resulting X/R ratio is traditionally ignored in the final computation. In this final step, after motor and source admittances are added, a traditional multiplier of 1.5 (for maximum phase, or the equivalent 1.25 for average phase) is used for d-c offset; this implies an assumed weighted-average power factor for source and motors of 8 per cent, or an X/R of 12.5.

As illustrated by the authors' numerical examples, this traditionally assumed X/R of 12.5 is high and therefore conservative. The authors' method reduces this conservatism and therefore results in a more economical application of switchgear.

However, it appears that the authors' method still retains elements of conservatism which should be reviewed before applying the curves of Fig. 2 to 9:

1. Primary source capability is assumed to be 500 mva. Many times it is actually as low as 30 mva.

2. X/R of the source, as seen from the lv bus, may be reduced if the transformer is cable-connected to the bus.

3. X/R of motors at locked rotor is usually half or less of the value of six assumed, for the size involved on most lv systems, and as seen from the bus is further reduced by cable connections.

4. For an a-c time constant of 2 cycles (range: 1 to 3½ cycles for lv motors[1]) the a-c contribution of an induction motor would decrease to 78 per cent in the first half-cycle.

When motor contribution forms an appreciable part of total short-circuit current (i.e., low-mva primary source), and if substation load consists predominantly of a few large motors, recognition of these factors will often permit switch-gear economies in borderline cases, provided allowance is made for the effect on short-circuit level of future expansion.

Two other assumptions made by the authors relative to motor contribution deserve comment. First, the assumed motor equivalent impedance of 25.3 per cent may be high. If the bus is heavily loaded and motors are of designs with high starting current, typical of totally enclosed fan-cooled designs, equivalent impedance may be as low as 12 or 14 per cent. Second, the use of Fig. 10 recommended by the authors is based on three questionable premises:

1. That prefault and postfault motor and source currents are in phase, so that the same degree of asymmetry or initial d-c contribution would occur in each.

2. That the d-c decay of the source and motors is equal, based on a weighted average X/R, and is not affected by the interaction of the two flowing through common circuit elements such as feeder cables.

3. That theoretical maximum initial d-c contribution from motors is equivalent to a multiplier of 1.73 for a fully offset wave; it would appear that this would occur only at no load and that the postfault current of a loaded motor, with a prefault current ⅛ to ¼ of symmetrical postfault, could be more than fully offset by an amount up to ⅛ to ¼, depending on prefault power and postfault X/R as seen from the motor.

Fig. 10 is captioned "Multiplying factors to obtain short-circuit asymmetrical current from symmetrical values, at an instant ½ cycle after initiation of a fault." It is wondered whether the curves are based on the decrement in d-c component reached after a full half-cycle, or whether they are for the decay in the first quarter-cycle and thus represent the average over the first half-cycle.

REFERENCE

1. TIME VARIATION OF INDUSTRIAL SYSTEM SHORT-CIRCUIT CURRENTS AND INDUCTION MOTOR CONTRIBUTIONS, W. C. Huening, Jr. *AIEE Transactions*, vol. 74, pt. II, May 1955, pp. 90–101.

W. C. Huening, Jr. (General Electric Company, Schenectady, N. Y., on behalf of Joint Sections Committee on Air Circuit Breakers of NEMA): We wish to express our appreciation to the discussers. Both are apparently interested in seeing how the material presented might be useful to them. Both have mentioned that there are particular cases where the short-circuit duties are lower than those given in the paper. This is true, but it is also true that there are other cases where the possible short-circuit duties are equal to those given. It is the purpose of the paper to present the maximum probable short-circuit currents for lv systems; therefore, it is not surprising that certain systems may have lower duties.

Engineers of the caliber of Mr. Clark and Mr. Kelly are quite capable of computing short-circuit duties in particular systems. Although it was hoped that such engineers might find the graphs useful for reference, the results were not compiled originally with that in mind. Instead, it was intended that the material should serve as the basis for a simple circuit-breaker selector guide which would be useful in determining applicable sizes of circuit breakers. It was expected that the guide would provide a way for men not skilled in the details of short-circuit calculations to select adequate circuit breaker equipment.

Mr. Clark will probably agree that each

power-plant-station service lv switchboard should be checked individually for short-circuit duty before circuit breaker ratings are selected. Average curves would probably be only of passing interest.

Mr. Kelly mentions that the method of calculation used for the study eliminates some of the conservatism of the traditional simplified calculating procedures. It should be noted that this introduces complications, and the advantage of simplicity is lost. The traditional methods are still the easiest way to calculate an answer that will permit the selection of an adequate circuit breaker.

Mr. Kelly also lists some ways in which the results obtained might be excessively conservative. This listing might be useful to an engineer interested in checking whether his particular system has lower duties than those given in the paper.

There is an area, as noted by Mr. Kelly, where the method of the paper may be non-conservative. There are possible cases where motor contributions might be greater than those computed. Mr. Kelly mentions two such situations: when the substation load consists predominantly of a few large motors, and when the load includes many motors that have high starting currents, typical of totally enclosed fan-cooled designs. It was decided not to treat cases like these when the method of calculation was selected, because overemphasized motor contributions would probably not be suitable for a general survey. It was considered unlikely that extreme types of motor loads would be found in systems where circuit breakers were selected by means of a simplified selector guide.

Some additional observations on the calculation of motor contributions might be made. The method of the paper employs a motor reactance higher than that corresponding to locked rotor current. Use of this higher reactance for the motor makes the effective motor short-circuit current contribution lower than locked rotor current. This lower current is intended to represent a motor contribution a-c component after it has decayed for ½ cycle from its initial value, which is theoretically equal to locked rotor current.

The use of an increased equivalent motor reactance introduces a slight error into the calculations when the point of short circuit is separated by appreciable cable impedance from the lv bus, and makes the results slightly inaccurate when multipliers taken from Fig. 10 are employed. The inaccuracies introduced are not appreciable. Other approximation having the same order of inaccuracy are also involved, not only in this but in other methods of making short-circuit calculations.

To comment on the "questionable" premises for Fig. 10, as listed by Mr. Kelly, it may be noted that prefault currents need not be considered in calculating short-circuit currents by methods such as the one used (unless saturation effects are to be included), and that any error introduced as described in Mr. Kelly's second premise is negligible until after the first cycle of short-circuit current flow. The short-circuit currents of the paper are first cycle rms values.

The multiplying factors of Fig. 10 do correspond to an average over one full cycle.

As a supplement to the paper, and aside from commenting on the discussion, mention should be made of other ways in which the results presented in the paper can be useful. Although marked for "liquid-filled" transformers, each curve should also apply to dry-type transformers, because they have similar impedances. Furthermore, each of the separate curves may also be used for other numbers of the same size conductors per phase, by employing the principle that a length of feeder with one conductor per phase has the same impedance as have two feeders in parallel each of which twice as long.

Cleveland Transit System Gets New Rapid Transit Cars

C. A. KOCH
ASSOCIATE MEMBER AIEE

T. H. MURPHY
MEMBER AIEE

CLEVELAND'S problems of transportation parallel, to some extent, those of her larger sister cities in the United States. The rapid growth of the population, the shift of population to suburbs, and the concentration of offices and trades in the downtown area made a new approach to the passenger transportation problem necessary. Early public transportation in Cleveland was almost entirely by surface street cars. This gradually was changed to a combination of street cars and buses, and finally led to a combination of trolley coaches and buses. In Cleveland there is now one single rapid transit line from the downtown section to Shaker Heights, operated by the City of Shaker Heights. It runs on its own right-of-way, using street cars equipped for multiple unit service. It is of more than national significance that Cleveland has recently inaugurated the first part of a rapid transit line which is designed to provide mass transportation, from east to west, parallel to Lake Erie. The part which is in operation now runs from E. 142nd Street, the Windermere Station in eastern Cleveland, to the Public Square in Cleveland.

General

Cleveland's is the third installation in the United States of a rapid transit line, with President's Conference Committee type of light-weight cars. The other installations are in Chicago and in the East Boston Tunnel. The Philadelphia and New York rapid transit cars are of a different type. The complete installation will cost in the vicinity of $60,000,000.

The rapid transit line is centered at the Public Square in the Cleveland Union Terminal Building. It runs east to E. 142nd Street through six stops, a distance of 8 miles from the Square with a scheduled running time of 18 minutes. A branch line is contemplated along Euclid Heights Boulevard later. The western branch is to run from the Square to W. 117th Street, a distance of about 5 miles with three intermediate stops and a schedule of 12 minutes for the run. An extension of approximately 2 miles beyond W. 117th Street is planned. The line is laid out to connect with trolley coach and self-propelled bus lines at every stop and, thereby, forms the backbone of an integrated transit system. New facilities were installed including tracks, signals, rectifier substations, catenary overhead power lines, and car maintenance shops.

Car Equipment

So far, 68 cars have been purchased. An effort was made to obtain light weight by the use of Cor-ten steel in all structural members of the body and by the use of a new light-weight truck and by special arrangement of the equipment. Significant dimensions of the cars are as follows:

Car length over end sills—single units, 48 feet 6 inches.
Car length over end sills—double units, 97 feet 2 inches.
Car width, 10 feet.
Car width at doors, not over 10 feet 4 inches.
Car height (pantograph in the down position), 13 feet 3 inches.
Car height (rail to roof), 11 feet 9 inches.
Truck center distance, 31 feet 2 inches.
Truck wheel base, 6 feet 6 inches.
Wheel diameter, 28 inches.
Door openings, 4 double doors, 2 on each side.
Door opening width, 4 feet 2 inches.
Door opening height, 6 feet 3 inches.
Seating capacity, 55 (single car).
Car weight—single units, 55,500.
Car weight—double units, 106,200.

The cars have a motorman's cab at each end of the single units, but only at the extreme ends of the semipermanently coupled units. Headlights and train accessories are provided in the same manner. Car bodies are insulated against temperature and sound. A smooth exterior contour permits easy cleaning of the cars. Silver-gray and blue colors were chosen as an attractive interior and exterior color scheme.

Trains

Cars are designed in two types—single and double unit. There are 12 cars arranged as single units and 56 cars arranged in 28 double units. The double units are connected semipermanently by a noncoupling draw bar. This arrangement of single and double units facilitates the adjustment of the train size to suit traffic conditions. The scheduled speeds, exclusive of layover, average approximately 28 miles per hour, with top speeds over 40 miles per hour.

Passenger Comfort

Special attention was given to passenger comfort and eye appeal in the interior, door arrangements, and quietness of operation. Platform loading is used at every station. The seats are fixed and face the nearest exit door. A seated passenger is never over 11 feet from the nearest exit. The seats are upholstered with flame-resistant fabric on foam rubber. Tubular stainless-steel bars are arranged across the tops to form grabhandles.

HEATING AND VENTILATION

The cars are equipped with an elaborate heating and ventilating system, thermostatically controlled, for maximum comfort. The panoramic windows are glazed with safety heat-resistant glass to keep out excess heat from the sunlight. There are four fans located symmetrically in the ceiling. These fans will deliver a total of 14,000 cubic feet of air per minute, taking air from the outside through a plenum chamber located in the roof. For control of the temperature, a system of dampers in the plenum chamber per-
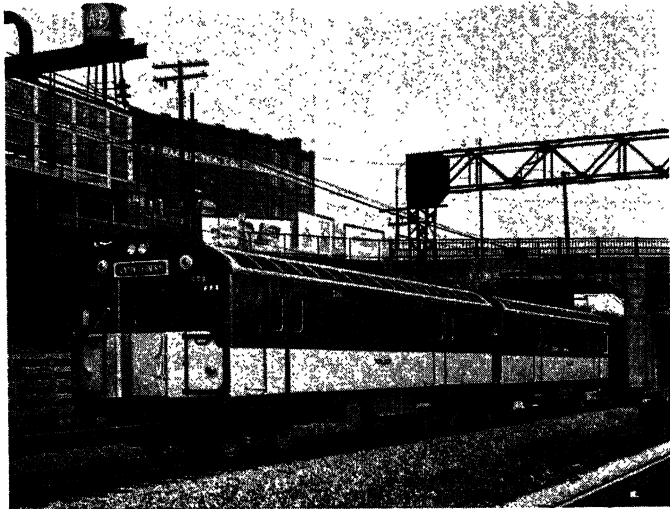
Fig. 1. Rush-hour service handled by two to six car trains

Fig. 2. The motorman's position, ceiling fans, fixed window, and seats

mits partial recirculation of the air. The fans start operating at 66 degrees Fahrenheit and reach their maximum speed at 85 degrees Fahrenheit. The car-heating equipment uses the heat from the accelerating and braking resistors. Two fans mounted on the underside of the car floor ventilate these resistors. They take the air from the car plenum chamber down through a structural post of the car. Additional auxiliary heaters are located in this air stream. If the temperature drops below 50 degrees Fahren-

heit, these auxiliary heaters are energized. The auxiliary heaters are cut off above 58 degrees and modulated heat is used up to 65 degrees Fahrenheit. Above 65 degrees Fahrenheit, the cooling air for the resistors is all passed to the outside.

LIGHTING

Lighting is furnished by fluorescent tubes mounted end to end in two diffusion-type channels at the ceiling, one row on each side. The intensity of the lighting is 20 foot-candles minimum at the reading

level. Emergency lights operated from the battery are located inside the fluorescent light fixtures and at the doors. Destination signs and marker lights are also provided at the car sides and front.

SAFETY

The four double doors are interlocked with the motorman's controls so that the train cannot be started if any door remains open. A signal light in the motorman's cab indicates when all doors are closed.

Fig. 3. Main power schematic. Acceleration and braking power circuits

SYMBOLS

LB = LINE BREAKER

①②③④ = MAIN MOTORS

LOR = LOCKOUT RELAY

LR = ACCELERATING & BRAKING RELAY (LIMIT RELAY)

$\frac{MB2-MB14}{MB1-MB12}$ = 135 POINT MOTOR OPERATED RESISTOR

PBC = POWER BRAKE CHANGEOVER SWITCH

Fig. 6. Control apparatus panel mounted on the underside of the floor
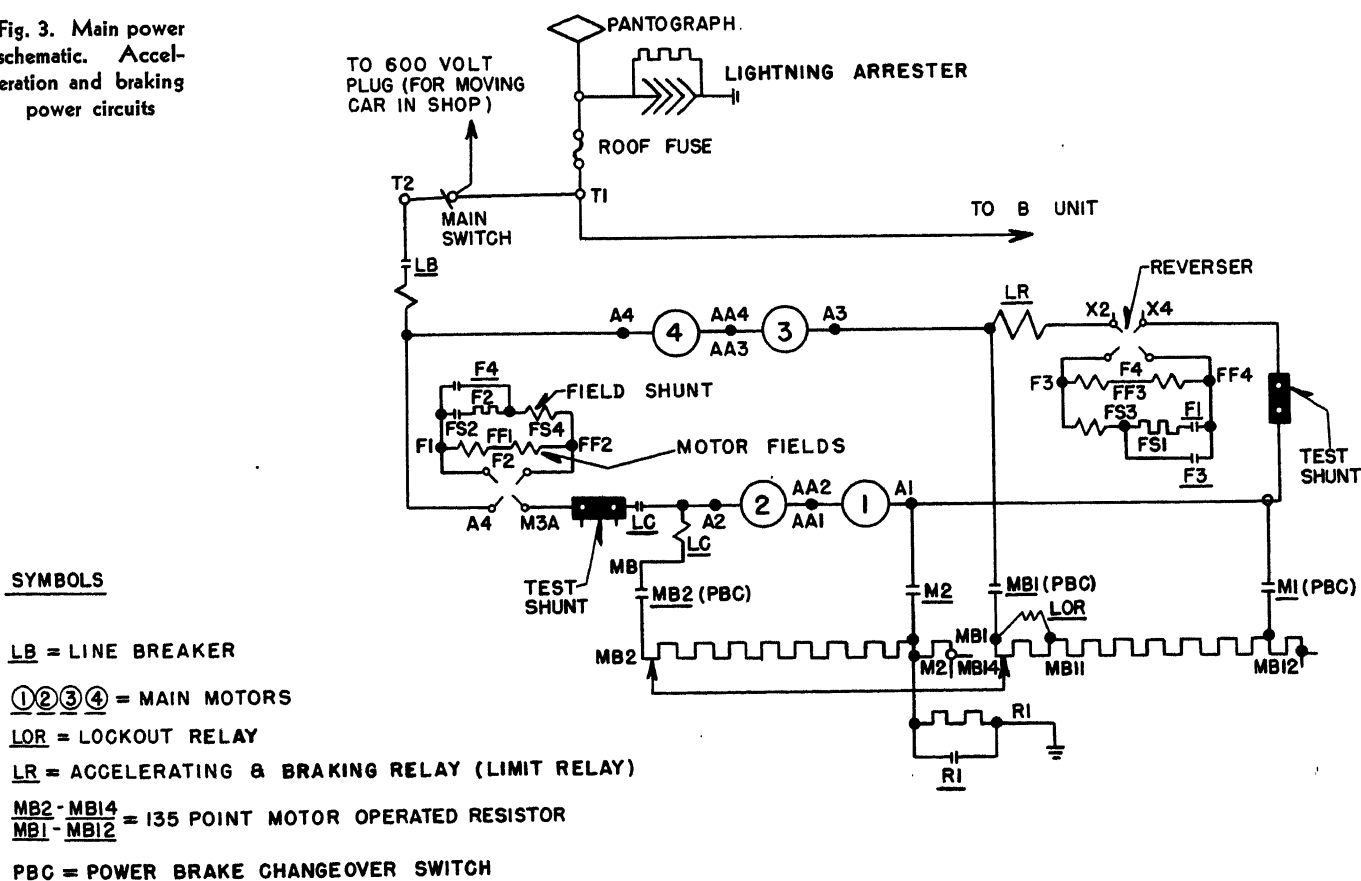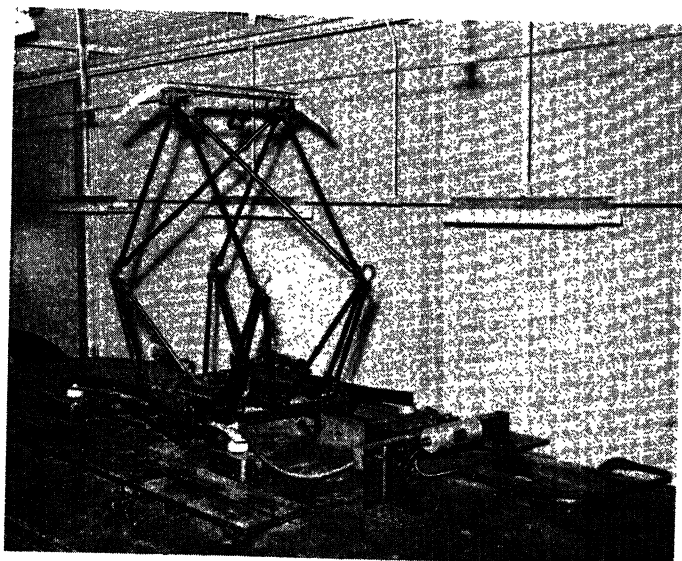
Fig. 4. Lightweight pantograph, fuse, and lightning arrester

Automatic trips are provided at each color light signal which engage a braking control lever on the forward car and bring the train to a stop in case the motorman should pass a red signal. A "dead man's" control stops the train automatically if the motorman should fail to hold down the control lever on the master power controller.

QUIETNESS OF OPERATION

The floor covering is marbleized black rubber in the aisles and dark blue under the seats on plywood. The equipment itself is designed for quiet operation by using automotive-type hypoid gears between the motors and axles, and four shock absorbers on each truck—two of which control lateral movement and two of which control lengthwise movement. Comfort of passengers during acceleration and braking is obtained by making these functions fully automatic and by producing accelerating and braking rates which

do not disturb the feeling of security of the passengers.

The motorman has a change window adjacent to the fare box for use during hours of light traffic when fare are not paid at the station. The station platforms are well lighted by continuous rows of fluorescent lights and escalators are or will be installed in major stations.

Electric Equipment

All axles are motorized, each by a 55-horsepower motor. The motors are connected two in series at all times. Acceleration is provided by an accelerating resistor with constant-current control. At higher speeds, the motor fields are shunted in two steps. The motors are self-ventilated by a shaft-mounted fan which draws air from the car body through duct work with flexible connec-

tions to the car body. The body air intake is located at the end of the car, alongside the sliding door. There is an air duct at each end of the car, serving one pair of motors or one truck. The two motors are mounted longitudinally on each side of the bolster, securely supported by the cross members. They are connected to the hypoid gear unit through universal joints. This contributes also to quiet operation and smooth riding because the motor weight is not carried directly by the axles. The motors are equipped with removable covers on the commutator end for ease of inspection of the commutator and for brush replacements.

Power circuits are remotely controlled by either magnetic or electropneumatic means. The control wiring itself is a 2-wire system operated on 40 volts d-c. Control of the train is centered at the motorman's cab. These cabs are equipped with a sliding window, convertible
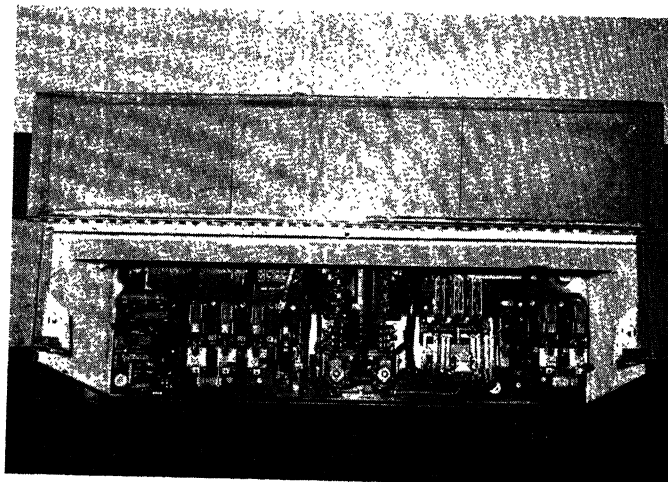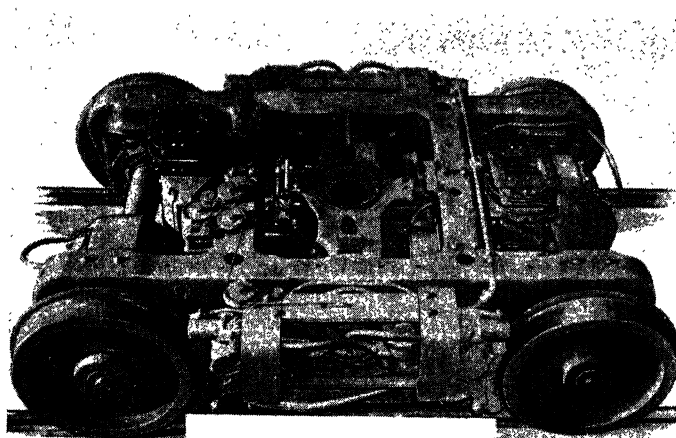
Fig. 5. A new design of a truck with two 55-horsepower motors
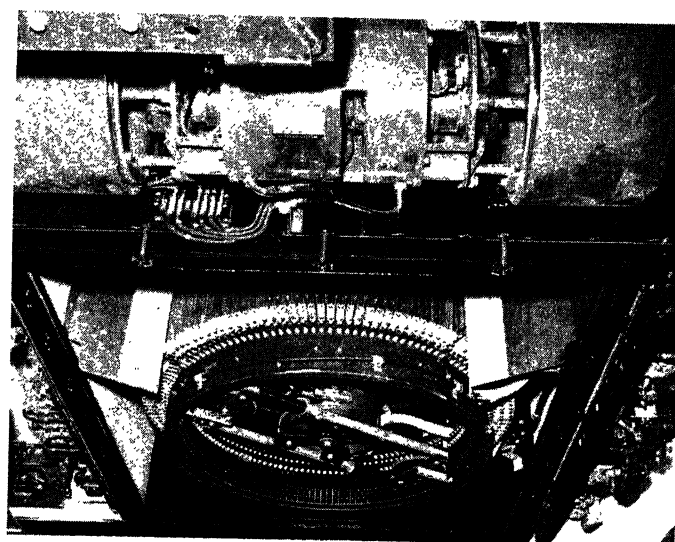
Fig. 7. 135-point acceleration resistor controller and blower set

seat, polaroid visor, electric heaters, glass windows on all sides, and a change window to the car interior. Other accessories include an air windshield wiper, a defroster fan, pantograph control buttons, door control buttons, sander and horn control, a fire extinguisher, and a master power and brake controller.

Acceleration and braking are automatically controlled through the master controller to provide uniformity and passenger comfort. The accelerator is operated by a fractional horsepower motor under the influence of a current-limiting relay which operates both in the advance and return directions. Acceleration is obtained by moving the master-controller handle to the left. Acceleration rates from 1 to 2.3 miles per hour per second may be selected. Coasting is produced by moving the master-controller handle to the central position. Moving the master-controller handle to the right sets up braking which starts with dynamic braking. Dynamic braking fades out at about 4 miles per hour where the air brakes take over the braking function. Braking rates range from 2.8 miles per hour per second to a minimum which occurs with the controller handle in the coasting position. Emergency braking is available when the controller handle is moved to the extreme right position and provides both dynamic and air brakes simultaneously. The master controller has a separate handle to operate the reverser. All cars are equipped with a load-weighing control device which recalibrates acceleration and deceleration depending on the load in the car so as to provide uniform acceleration and deceleration rates regardless of loading conditions.

Power is taken from the overhead line by pantographs, one for each single unit and one for each double unit. The pantograph on a double unit is on the forward car. The pantographs are mounted on the roof at the center of a truck. They are of light tubular construction with single contact shoes, spring-raised and air-lowered. The single shoes have a pan extending the width of the framework with lightweight aluminum horns on the ends. The pan carries two strips of copper for its full length. Gra-

phite lubricant is carried in the pan between the copper strips to minimize wire and copper strip wear. The working range of the pantograph is 4 feet. Operation is effected by two extra long coil springs which give a uniform working pressure over the operating range. Two air pistons mounted on the base are used either to lower the pantograph against the action of the springs or to release a mechanical latch which holds the pantograph in the down position. If air pressure is not available, the latch can be released by a pull cord. Maximum current per pantograph is approximately 1,000 amperes for a 2-car train including lighting and heating load. Pantographs are protected by a fuse against overload. Lightning arresters are installed on the roof adjacent to the pantographs for protection against lightning currents.

Auxiliary power for the control functions is furnished by a 40-volt battery system. The battery is charged by a generator from an auxiliary motor-generator set. Main control panels and auxiliary control panels are mounted on the underside of the car with sliding or hinged doors for easy access during maintenance and inspection.

Wiring from car to car is carried by control receptacles on the single units which are incorporated in the couplers, and by a junction box with flexible jumpers between the double units. The 600-volt power circuit jumpers and air lines between the cars of a double unit are clamped to the draw bar.

Mechanical Equipment

Compressed air for operation of the air brakes and auxiliary devices is furnished by a 2-stage motor-driven air compressor of 25-cubic-foot-per-minute capacity. This compressor is also mounted on the underside of the car floor. It supplies air pressure to operate the brakes. Each truck has one cylinder and one brake shoe per wheel. This results in a very simple brake rigging. The compressed air is also used for operating some of the contactors and for operating the doors.

New light-weight trucks were designed for these cars. The trucks are of the inside frame pedestal type, cast steel, with

tapered roller bearings and a spring-supported bolster. As mentioned previously, a silent right-angle hypoid gear drive and the use of shock absorbers contribute materially to riding comfort.

Maintenance and Shop Facilities

Shop facilities have been provided at Windermere for complete maintenance and inspection operations. One inspection pit has been equipped with eight hydraulic jacks so that the two bodies of a double unit can be supported simultaneously for truck removal. The four trucks then can be lowered to a working area one floor below the track level. In this manner, the cars can be returned to service with a minimum delay. In addition, tracks are provided over pits to service and inspect two double units. These pits are 4 feet deep and extend $2\frac{1}{2}$ feet beyond the side of the rails so that a quick inspection can be made of the apparatus panels which face toward the sides of the car.

Conclusions

Cleveland is providing a rapid transit system which will restore the downtown area as the center of all activities in commerce and permit planning on a large scale for the future of the downtown area. The rapid transit line will carry passengers quickly, efficiently, and comfortably. It forms the backbone of a coordinated transit system with feeder lines to the rapid transit stations and transfer privileges from and to the rapid transit line. Parking areas will make the rapid transit system also available to individual car drivers.

The present rapid transit operation is only the first step in a complete rapid transit system. The western portion of the line also will soon be in operation. The downtown loop which is a part of the original plan cannot be completed for several years. Only when this portion is finished can the full benefits be obtained. There is no doubt that Cleveland's rapid transit system will achieve the same degree of success which is characteristic of these developments in other large cities.

Operating Experience With a Mechanical Rectifier

J. CHAMULAK
NONMEMBER AIEE

W. C. McCULLOUGH
NONMEMBER AIEE

J. W. TRACHT
MEMBER AIEE

THE first 10,000-ampere 60-cycle mechanical rectifier installed in the United States has now completed 32 months of operation. Anticipated high conversion efficiency has been confirmed. The current supplied the isolated electrolytic load has been controlled well within the required limits. Contact life remains the major maintenance problem. However, recent developments in contact design have increased contact life and give promise of still further increase.

In May 1951 a 10,000-ampere 267-volt d-c mechanical rectifier manufactured by the I-T-E Circuit Breaker Company was placed into service at the Wyandotte, Mich., plant of the Pennsylvania Salt Manufacturing Company.

Description of Equipment

The rectifier unit, controls, d-c switchgear, and the air conditioning unit were installed in a separate building attached to the main cell building, with the transformer located out of doors on a concrete platform. See Fig. 1. The a-c oil circuit breaker is located inside the power plant approximately 200 feet from the rectifier. Inverse time overcurrent relays with instantaneous attachment are installed on this 8-cycle breaker.

The power transformer is Askarel-filled, self-cooled, sealed, and rated for a 45-degree-centigrade temperature rise. Location of the transformer near installations representing high investment dictated the selection of a noninflammable insulating liquid. Two $2^1/_2$-per-cent no-load taps above and two $3^1/_2$-per-cent taps below the normal system voltage of 13,800 volts are available as required. The no-load tap changer is controlled by an operating handle in the rectifier compartment.

Voltage regulation is provided down 50 per cent of maximum rated voltage for starting purposes only and down 15 per cent of maximum voltage on each transformer tap for continuous operation.

Primary and secondary windings of the main transformer are concentrically wound. The secondary is wound between two halves of the primary to keep the reactance to a minimum and the pri-

mary taps are at the center of each phase. No surge protection is needed for mechanical rectifiers. The interphase transformer is located outside the transformer tank on the d-c side of the rectifier and it is air-cooled.

The commutating reactors are mounted in the same tank with the transformer windings. Three control windings are used on the commutating reactors. One winding. that on the "break" core, completes the flux change in the reactor after the contact opens so that the recovery voltage on the contact is delayed until the contact is separated sufficiently to withstand it. This winding is connected to the break pre-excitation supply.

One winding on the "make" core compensates for the magnetizing current in the reactor by providing exciting current to it before the contact closes so that there is negligible inrush current at the instant of contact closure. The second winding on the make core is used to prevent the parallel resistor circuit from operating during the make step. This resistor circuit acts as a load on the reactor during the break step to prevent inductive peaks on the contact as it is opened.

The rectifier d-c breakers are of the high-speed type to permit future parallel operation with other power conversion equipment. The d-c breakers are single-pole, high-speed, reverse-current trip, type-*FB* circuit breakers rated at 6,000 amperes, 750 volts d-c maximum. Tripping by reverse current limits the current rise within 6.25×10^{-3} seconds. Complete interruption of the current occurs within 12.5×10^{-3} seconds. There is no maximum interrupting capacity established for the *FB* breaker. The latching mechanism consists of a spring-loaded holding magnet which provides the means for reverse current tripping when reversal of flux in the magnet releases the latch. A shunt trip and an emergency hand trip are also provided.

The a-c and d-c short circuiters are actuated by shifting the flux of a permanent magnet which latches the short circuiters in the open position. A flux-shifting coil is wound on the magnet and this coil is energized by the impulse transformer when an unbalance occurs between the

a-c and d-c systems of the rectifier. The closing time is 1.25×10^{-3} seconds from initiation of the pulse. See Fig. 2.

Controls may be divided into three groups: startup and shutdown sequence, load, and protective devices.

Startup and Shutdown Sequence. These are automatically controlled by electrically operated relays and starters. Startup is initiated by the a-c breaker control switch which closes the a-c breaker and energizes the rectifier transformer. Oil and water pump motor, regulator motor, and exhaust fan motor starters close and, when the oil pressure switch closes, the drive motor starter is closed. Voltage is applied to the rectifier after a time delay by starters, first through a resistor circuit and then directly. The motor-driven secondary disconnect is closed after the full secondary voltage has been applied to the rectifier directly through this starter stage. The automatic sequence stops here. Load is connected by closing the hand-operated d-c disconnects in the negative bus and then operating the d-c breakers electrically by a control switch. Shutdown procedure is the reverse of startup and it is automatic after the d-c breakers have been opened by operation of this control switch and the a-c breaker control switch is turned off. After the transformer secondary voltage is disconnected from the rectifier in the secondary resistance stage, the a-c breaker opens and the unit is shut down.

Load. No-load (base load) voltage control and load current control are achieved by the hydraulic regulator which is sensitive to voltage before the d-c breakers are closed and to current after the load is connected. The hydraulic regulator rotates the drive motor stator to maintain a fixed voltage or current which is determined by hand-operated rheostat controls.

Protective Devices. Immediate tripping of the a-c breaker is effected by short-circuiter tripping, transformer tap changer latch movement, instantaneous overload relay sin the transformer primary circuits, or emergency trip button. The overload relays are combination time delay and
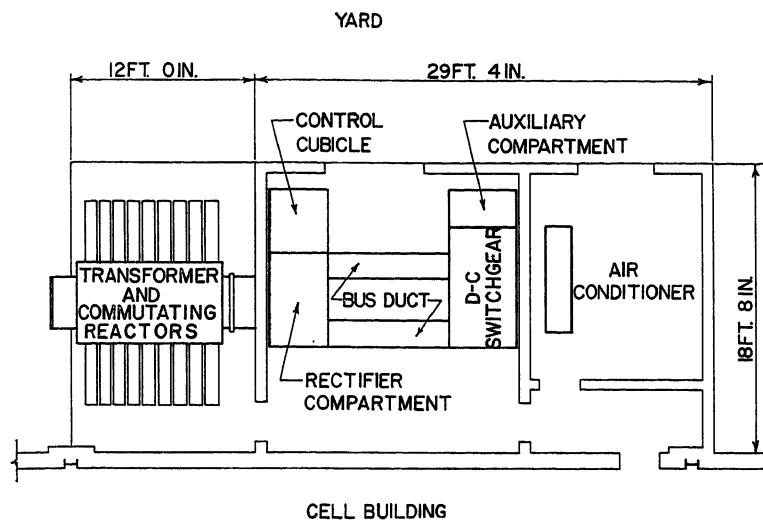
Fig. 1 (above). Rectifier station general layout
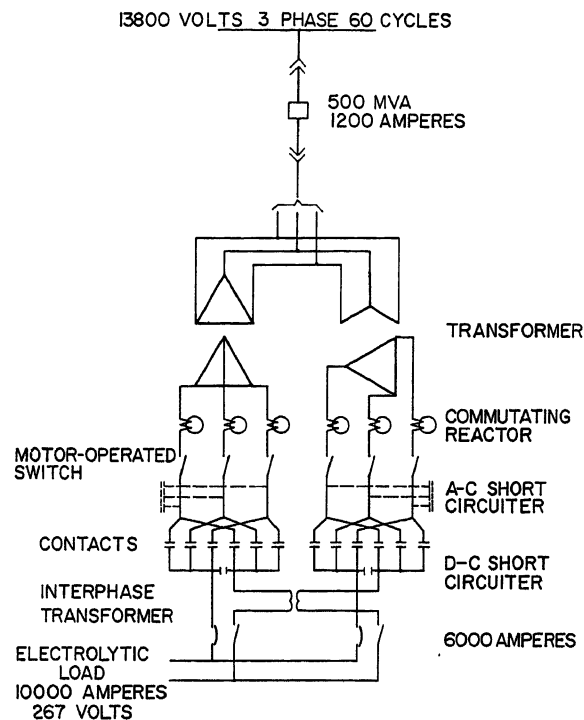
Fig. 2 (right). Rectifier basic circuit diagram

instantaneous. Normal sequence shutdown including automatic opening of the d-c breakers is effected by a master monitoring relay which is tripped if any motor starter fails to operate or opens while the rectifier is in operation. An alarm system, including a bell and red lights, is energized if transformer pressure or temperature exceed rated limits, if the transformer secondary load is greater than 1.5 times rated for a time in excess of 1 minute, or if the oil pressure falls below a low limit. The alarm system operates also on any fault shutdown.

Two continuous strip-recording type of instruments are flush mounted on the door of the auxiliary compartment. One instrument records the total direct current output. The second unit records and integrates the total d-c power output of the rectifier. Total current values to these instruments are measured from a common shunt in the main bus run to the load.

Automatic control is incorporated to maintain direct current output within ∓ 1 per cent. Setting of current to different values is accomplished by adjusting the position of a rheostat mounted on the control cubicle. Infinite step current control is provided at each setting.

Initially, a water spray type of air washer was installed to reduce the amount of corrosive fumes in the air and to provide cooling air for the rectifier equipment. Under certain circumstances, however, corrosive gases from nearby process areas were pulled through the air washer, mixed with the highly humid air, and circulated through the rectifier room. Rapid corrosion took place on exposed metal parts such as relay contacts, wire

terminals, rheostats, etc. To prevent possible damage to a major piece of equipment by improper relaying or improper operation of controls caused by this corrosion, the air-washing equipment was replaced by a commercial air-conditioning unit operated in a closed system thereby eliminating the need for make-up air.

The current-carrying portion of the contacts for the machine are of a silver alloy with the movable part spring-loaded and were originally unguided. These component parts are mounted in a plastic body. The contacts are attached to the a-c and d-c bus bars and thus are cooled by the liquid cooling system serving these bus bars. This is a closed system with a water-to-air heat exchanger. Opening of a contact is accomplished through movement of a vertical push rod with an insulated head driven by a cam arrangement.

A-c power for the rectifier is obtained from the plant generation system bus. This 46,000-kva system includes two 9,375-kva generators paralleled on a 13,800-volt bus and five generators, ranging in size from 3,125 kva to 7,500 kva, operated at 2,300 volts in a synchronizing bus arrangement. The 13,800-volt system is tied through transformers to bus sections of the 2,300-volt system. Synchronous converters are operated on both systems and paralleled on the d-c side. This places the rectifier in parallel operation on the a-c side with synchronous rotary converters and mercury arc rectifiers delivering d-c power to electrolytic

loads. To date the mechanical rectifier has not been operated in parallel with other conversion equipment on the d-c side. A manually operated throwover switch has been provided to transfer the load served by the mechanical rectifier to the main d-c bus in event of a rectifier shutdown.

When placing the rectifier in service, individual units of the electrolytic load are connected to the rectifier one at a time. Under this procedure, the rectifier output current is regulated to the value required for that portion of connected load and then is raised to approximately meet the load demand of the next incoming unit. By following this practice, the value of inrush current to the rectifier is limited to a maximum of 15,000 amperes.

Contact Life

Since this rectifier was the first 10,000-ampere unit to be operated on 60 cycles in the United States, contact life was an unknown factor. It was necessary, therefore, during the early stages of the rectifier operation to keep the rectifier running until one or more contacts failed to obtain contact performance data. The program of studying the behavior of contacts consisted of noting the condition of each contact during the break step portion of the commutating period as instructed by the manufacturer. By means of a selector switch, the voltage wave shape of the break step of each contact was observed hourly on the cathode-ray oscilloscope and the condition re-

corded. This condition was expressed as "good," "bouncing or irregular break," or an "arcing" contact by comparison with sample voltage wave drawings supplied by the manufacturer. Since these checks were made every hour, the interpretation of the oscilloscope picture as to contact condition was done by the process operators.

After compiling records for 3 months, an approximate average life of a set of contacts had been determined and the program was discontinued. However, contact failure still could not be accurately anticipated. Contacts that appeared to be in good condition would change to show signs of a poor condition, run this way for a number of hours and then show a good condition again. At other times, contacts that appeared to be operating favorably would suddenly fail. Contact failure is considered to have occurred when the surfaces of the contact have roughened to the extent that continued operation is not possible.

Because of the uncertain remaining life of the contacts of a set in which failure of several contacts has occurred, it was found desirable to initiate a practice of replacing all contacts after any were damaged. From the accumulated records an average operating life of 300 hours between backfires for a set of contacts was established. It was then decided to institute a program of replacing all contacts after 275 hours of operation.

Procedure for Contact Replacement

To replace a complete set of 12 contacts and return the rectifier to service requires an average of approximately 3 hours. Work is accomplished by the best electricians in the plant normally during the regular day shift. Should the rectifier trip out of service at any other time, contact replacement is postponed until the next day shift. The process operators perform the job of transferring the rectifier load to the main d-c bus to maintain production.

No attempt has been made to decrease the time of contact changing because of the available stand-by power for the cell load, thus allowing time for proper training of all qualified electricians, general inspection of control equipment, and checking the progress of electrolytic attack on the cooling-water system. Any other required maintenance is performed during the downtime of the rectifier for contact replacement.

Contact replacement procedure begins by loosening 18 single bolt bar clamps to remove the old contacts. Each contact is marked as to the time installed and the time removed, using readings from the time meter provided. The position of the contact on the mechanism also is recorded. This information is forwarded to the manufacturer to assist in the program for improved contact life.

The top of the contact mechanism is cleaned of all minute particles of contact material that were deposited when the backfire occurred. Loose pieces of contact material in countersunk screw holes in the push-rod heads are removed to eliminate the possibility of a piece being caught between the push-rod head and the movable contact during normal operation. Should this happen, it could affect the timing of the contact.

Now the new set of contacts is clamped in place and the contact drive motor is energized from an external source of auxiliary power for the initial adjustment of contact timing. The auxiliary power circuit is provided in the control cubicle. This timing is accomplished by rotating an external screw geared to each individual push-rod. The adjustment is set to obtain a predetermined steady direct voltage across the contacts. Most of the time consumed in replacing contacts is at this stage. Usually two to four contacts are difficult to time properly. At times, by driving the contacts for a short break-in period, the poor timing will clear up. Should this fail, the contact will need to be replaced. Running the contacts at no load for a 30-minute period or longer after initial timing is beneficial because much of the contact flaking takes place at this time. After all the contacts have been properly timed by the foregoing voltmeter method, the rectifier is shut down and the auxiliary power disconnected.

The a-c circuit breaker is then closed and the rectifier components are placed in operation by automatic selective relaying. Provision is made, however, to interrupt this automatic sequence for checking contact operation before continuing the automatic starting. By means of manual-automatic selector switches, a fused starting resistor is shunted across the secondary disconnect switch on all phases and in series with the contacts. The operation of each contact is then viewed in turn on the oscilloscope to determine whether it is opening within the break step. This operation requires only a matter of seconds for each contact.

A quick check of the contacts is completed using the oscilloscope for timing of mechanical break, correct amount of pre-excitation voltage, and contact opera- tion. When all contact conditions are satisfactory, the two contactors are placed on automatic position, thereby removing the starting resistors from the circuit and permitting the secondary contactors to close in proper sequence for continuous operation of the rectifier. At this time the unit is available for service.

With the machine running on base load, the process operators assume control of the rectifier and perform all switching operations during the application of electrolytic load to the rectifier. The voltage control rheostat is adjusted to provide proper voltage output of the rectifier for application of the first unit of load. The current control rheostat is turned back to a point that corresponds to the amperage drawn by this unit. Once the rectifier d-c breakers are closed, the voltage controller becomes ineffective and the current controller is used to load the rectifier as the load units are individually connected to the d-c output bus. All the available load is applied to the rectifier by the operators in a matter of 2 to 3 minutes. When this operation is completed, the electricians make final adjustments as required on the contacts. Some of the contacts that show a bouncing condition or signs of arcing are cleared of the difficulty by a slight change in time of mechanical break. This is successfully done while the contacts are carrying load.

An hourly check of the contacts is performed by the process operators during their regular process inspections and a record is made of any irregularity in contact performance. A check on the contacts once a day is maintained by a qualified electrician.

Efficiency

Calculations based upon the readings of the a-c induction watt-hour meter and the d-c potentiometer type of watt-hour meter with mechanical integrator have established an efficiency of the rectifier and transformer of 97 per cent when operated at 85-per-cent load. Accuracy of these two meters has been maintained within 1/2 of 1 per cent under this load condition.

Power Factor

Operation with minimum phase control is practiced to obtain the best power factor. With no phase control, power factor is 0.88 to 0.90, dependent upon the transformer tap being used. With phase control the power factor drops off directly as the percentage reduction in d-c output voltage.

Other Maintenance Problems

Leakage currents cause electrolysis in the cooling-water system. Various types of fittings are being experimented with for interconnecting the plastic hoses of the cooling system to the bus headers. The intent is to control the location of the points of electrolysis and thereby reduce time in changing the eroded fitting.

Design Improvements

Short nonuniform contact life is caused by the mechanical failure of the contact to maintain a predetermined length of closure time for each operation. Material transfer between contact surfaces causes surface irregularities. The closure time is altered if the matching surfaces do not mate in exactly the same position on each closing operation. If the contact remains closed too long, it is forced to open while carrying a reverse current which will destroy it.

Roughening is caused by mechanical friction at the contact surfaces. Bouncing, which results in light arcing, and discharge of the voltage across the contact as it approaches the closed position cause material transfer.

The contact material is a silver alloy because of the need for high conductivity since the contact area is small. When new, the contacts tend to flake heavily and there is evidence of galling in most contacts after about 100 hours of mechanical operation. Indium plating of the contacts shows a marked decrease in the tendency to flake or gall while very high conductivity is preserved.

Contact construction has been altered to provide a method for guiding the moving contact into the same position on the stationary contact on each closing operation. Material transfer is markedly reduced in this type of construction and, when irregular contact surfaces develop, the contact closure time is not affected in an amount sufficient to cause failures. The use of the new contact design on a limited scale on this unit promises periods of successful operation of at least 1½ times those experienced with the original contacts.

The use of water cooling for the bus structure upon which the contacts operate and the lubrication system have been discontinued by the manufacturer of this rectifier. Newer units are equipped with forced-air-cooling systems in which the air-circulating fan replaces the water pump and water heat exchanger and the lubricating oil is cooled by a small heat exchanger at the intake of the circulating fan. This design is more compact than the water-cooled one and a complete assembly of the air-cooled mechanical rectifier mechanism and bus work is interchangeable with the water-cooled type. Replacement of the present water-cooled unit in this rectifier with the air-cooled type is under way. Since the atmosphere in which this installation is located is well controlled, a considerable saving in maintenance costs because of elimination of repairs to the water-cooling system is expected.

Discussion

Otto Jensen (I-T-E Circuit Breaker Company, Philadelphia, Pa.): As of January 1954 the rectifier was in operation for 32 months. At this time the rectifier has now been operating 50 months. The change of the cooling system has been made and the expected improvement in maintenance has made in the contact construction and considerable longer life is now experienced with this machine.

The mechanical rectifier is constantly being improved and, as this device is a relatively new one compared with other conversion means, it is to be expected that improvements will be continued for some time to come. In the meantime, machines already delivered and currently being delivered can be looked upon as a practical means of supplying d-c power. The following improvements have taken place.

CIRCUITS

The so-called 6-coil connection in which a commutating reactor is connected in series with each contact of a 6-phase system is now universally used. This circuit permits closing of the contacts at voltage zero, at which time there is no voltage across the contact. This eliminates one annoying source of material transfer between the mating contact surfaces. The commutating reactors are now used for voltage or current regulation of the rectifier giving a voltage regulation of from 30 to 40 per cent in downward direction against the 15 per cent permissible by the mechanical phase shifts previously used.

On the low-voltage machines, i.e., in the range of 50 to 250 volts, the star connection of the rectifier transformer is used resulting in doubling the current output of the machine. Ten thousand amperes for a 6-contact machine and 20,000 amperes for a 12-contact machine can now be realized. This increase in the current-carrying capacity of the machine results in a lower overall cost to the customer.

The use of the 6-coil connection has raised the voltage level of the bridge-connected mechanical rectifier from the former 400-volt limitation to 1,000 volts d-c. In Europe one mechanical rectifier is now operating at 880 volts d-c.

CONTACTS

The contact life is still not predictable but improvements have been made. Some installations report contact life of more than 1 year, others are not so fortunate. It can, however, be said that better guiding means of contacts are now available which reduce the mechanical wear materially, thereby eliminating the flaking effect mentioned in the paper. Furthermore, better and easier mounting means have been provided permitting simpler ways of changing the contacts. Contacts in a 12-phase machine can be changed and the machine placed back in service in 10 to 15 minutes. It is not necessary to run the contacts in before placing the machine on load. Further work is being done in regard to contact life and we will not be satisfied until 6-month contact life is obtained as a minimum in the 6-coil connected machine.

PROTECTION

As the mechanical rectifier will backfire when the feeding a-c system is subjected to severe voltage dips, and perhaps also from internal trouble, it is necessary to protect the contacts from being destroyed by the resulting electric arc. This is done by furnishing a better path around the contact for the fault current. This path is provided by a short-circuiter, also called a contact protector. This device must be fast in its action as the damage done to the contact increases with the square of the operating speed of the short-circuiting device. Installation is now being made on an improved short-circuiter which is approximately 100 times faster than the short-circuiter it replaces. This increase in speed will result in so little damage being done to the contact during a backfire that a machine can be restarted without bothering to change the contacts.

In some installations consisting of more than one machine, sympathetic backfires occur if one of the machines gets into trouble. One installation consisting of three 5,000-ampere mechanical rectifiers has been equipped with the so-called d-c step reactor which prevents parallel machines from feeding into the faulted machine before its d-c breaker has been opened and cleared the defective machine from the bus. Therefore, the parallel machine is not severely overloaded on the d-c side and cannot backfire because of overloaded conditions. This is in contrast to the conditions which existed before the step reactor was installed when all three machines were invariably lost. It must, however, be pointed out that sympathetic backfires may also be experienced when the feeding a-c network is weak. In this case the backfire of one machine will depress the alternating voltage to the extent where sympathetic backfires will occur because of severe voltage dip on the a-c side.

SYNCHRONOUS DRIVE MOTOR

The exact timing of the contacts in the machine is obtained from the synchronous

drive motor. If this motor falls out of step or if a great phase shift occurs between the stator and rotor, the exact timing is lost and the machine will backfire. I-T-E has now completed tests on a motor where the motor will stay in step during severe alternating voltage dips and it is expected that when this motor is available it will materially reduce the number of unexplained shutdowns which have been experienced in the past.

The mechanical rectifier cannot as yet claim the connected capacity that the mercury arc rectifier enjoys but there was a time not so long ago when the mercury arc rectifier was considered new and unproven and it took a good many years for the mercury arc rectifier to get general acceptance in this country. The mechanical rectifier, however, is doing a creditable job

in the electrochemical industry. For the first time the user of d-c power does not have to employ high direct voltage to obtain high efficiency. Therefore, today low voltage and high currents can be used in such processes as production of hydrogen chlorine, aluminum, peroxide, and others. The soundness of the I-T-E pioneering work in this field is amply illustrated by the other manufacturers now joining us in our endeavors.

.

J. Chamulak, W. C. McCullough, and **J. W. Tracht:** Since the paper was first presented an additional 18 months' operating experience has been obtained, making a total of over 4 years' successful operation. Developments in contact design have substantially increased the contact life, which

now averages 800 hours. The unit has been operated up to full load with efficiency of 97 per cent still being realized. The problems of leakage current causing electrolysis in the cooling-water system have been eliminated by substitution of air cooling of the bus structure and lubricating oil systems.

Mr. Jensen's discussion very adequately summarizes the more recent developments in design of the mechanical rectifier which still further increased the range of application and improved the reliability of this high-efficiency equipment. It is encouraging to note the progress being made in combating the problem of sympathetic backfires in installations of more than one machine and in minimizing contact damage during backfires through the use of higher speed protectors.

Analysis and Design Principles of Second and Higher Order Saturating Servomechanisms

R. E. KALMAN
STUDENT MEMBER AIEE

Synopsis: This paper presents an extension and generalization of the phase-plane method to automatic control systems governed by high-order nonlinear differential equations. A unified procedure of analysis is outlined. It is based on linear transformations in the phase space, correlated with the partial-fraction expansion of transfer functions to separate natural frequencies, and makes use of the root-locus method for the qualitative study of closed-loop stability. The analysis leads to replacing a high-order system with a second-order system which closely approximates the former. Excellent insight is obtained into difficult problems. The method is applied to a study of control systems subject to saturation. Appendix II summarizes the current state of knowledge concerning second-order optimum saturating systems.

SINCE about 1950, there has been an awakening of serious interest in the problems of analysis and design of nonlinear automatic control systems, i.e., systems where the principle of superposition is not applicable. Already, considerable clarification of such problems has been possible by the use of the powerful phase-plane (topological) method.[1-16]

It is generally believed at present that the phase-plane method is suitable only for the analysis of second-order systems. As will be seen, however, the advantages

of the method may be retained even in the analysis of high-order systems, provided that simple means can be found for approximating in a plane the cardinal features of trajectories in n-dimensional phase space. The purpose of this paper is to outline such an approximation technique, and to illustrate it by numerous applications to the problem of saturation. This paper covers the second stage of a long-range study of nonlinear control systems.

Essentials of the Saturation Problem

The engineer is vitally and urgently interested in developing a good theory of nonlinear dynamic systems because he wants to understand fully the effects on system performance of the various inevitable physical limitations and imperfection of practical equipment. The most obvious limitation may be termed "saturation." It is everywhere present. For example, the voltages in a transformer are limited by magnetic saturation; the amplitudes of signals handled by electronic amplifiers are always finite; the flow-rate through a hydraulic or pneumatic valve, under constant pressure conditions, reaches limits when the valve is fully opened or closed; no one may

work for more than 24 or less than zero hours a day, etc. Thus saturation may be defined as an upper or lower limit or both on a physical quantity. In control systems, saturation in the output element will always limit the speed of response.

What is the physically realizable maximum speed of response of a system for a given level of saturation and how can optimum performance be attained or approximated? These questions are being answered by the recently developed optimum relay servomechanism theory discussed in Appendix II. The realization of optimally fast transient response, if at all possible, requires nonlinear compensation. Such compensation may present grave engineering difficulties, and it is important to know how closely the optimum can be approached by use of standard linear compensation methods. In this paper, the capabilities of some such methods (gain adjustment, tachometer feedback, cascaded lead or lag networks, etc.) will be explored in particular cases. This will be done so as to obtain as much insight as possible about various aspects of transient behavior without explicitly solving the nonlinear differential equations.

Principles of Analysis

The ensuing discussion rests on the following fundamental considerations:

1. To represent all solutions of an autonomous (time-invariant) ordinary nonlinear differential equation of the nth order, it is necessary to use an n-dimensional space, called the phase space. Each point in the phase space corresponds to a particular set of initial conditions and therefore determines a unique solution (trajectory) of the differential equation.

2. The trajectories of a nonlinear differential equation may be approximated with arbitrary accuracy by breaking up the phase space into a sufficiently large number of regions, such that within each region the trajectories obey a linear differential equation with constant coefficients. This is equivalent to approximating nonlinear characteristic curves by straight-line segments.[13] In particular, it will be convenient to describe saturation, from the start, by the idealized curve of Fig. 3.

3. To be able to deal with linear systems in n-dimensional phase space, as required by the scheme of the foregoing point 2, a further approximation is essential. It consists of replacing an nth-order linear system by a second-order one which accounts for the dominant (slowly-decaying) transient terms; the same idea is frequently used in linear servomechanism theory. If this approximation were carried out at the beginning of the analysis, all the interesting properties of high-order nonlinear systems which are not possessed by second-order systems would be lost. But if the approximation is postponed until the end of the analysis, the error committed will be very small and, as will be shown, the distinguishing features of high-order systems can be preserved.

It is necessary to develop a fairly elaborate mathematical apparatus for this second type of approximation. The central idea is the linear normal co-ordinate transformation, discussed in the following.

It should be emphasized that the results of this paper concerning the practical design of saturating control systems will be primarily qualitative, revealing phenomena but not usually supplying numerical answers. Such is the proper task of the phase-space method.[13] Even though all the phase-space analysis will be carried out in precise, quantitative terms, for the point-by-point calculation of transients various specialized graphical or numerical techniques may prove more suitable.[13]

Nomenclature

Properties and Representation of Linear Systems

The following will be the principal tools of analysis:

1. Theory of second-order nonlinear differential equations in the phase plane.[13,17-19] In particular, the terminology of reference 13 will be followed throughout and its contents assumed familiar. For convenience, a review of the phase-plane trajectories of standard second-order linear differential equations is given in Appendix I.

2. The root-locus method for studying the variation of the eigenvalues (roots of the characteristic equation) as a function of loop gain or other parameters.[20,21]

3. Vector representation of linear differential equations in the phase space.[22,23] A comprehensive discussion of this technique, interpreted with the aid of the signal-flow graphs, is the subject of the present section. Basic concepts of vector algebra, vector spaces, and linear transformations are treated from the modern point of view by Halmos[24] and Birkhoff and MacLane.[25]

STANDARD FORM RELATED TO SIGNAL-FLOW GRAPH

Every ordinary, nth-order, linear differential equation system with constant coefficients can be written in the standard form

$$\frac{dc_i}{dt} \equiv \dot{c}_i = \sum_{j=1}^{n} a_{ij}c_j + r_i \quad (i=1, 2, \ldots n)$$

or

$$\frac{d\mathbf{c}}{dt} \equiv \dot{\mathbf{c}} = A\mathbf{c} + \mathbf{r} \tag{1}$$

where $\mathbf{c} = \|c_1, c_2, \ldots c_n\|$ denotes the components of a vector, drawn from the origin, locating a point in relation to an orthogonal Cartesian co-ordinate system in the n-dimensional phase space $\mathcal{C}$;

$$A = \|a_{ij}\| = \begin{Vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{Vmatrix}$$

is an $n \times n$ matrix with real constant elements; the vector $\mathbf{r} = \|r_1, r_2, \ldots r_n\|$ denotes the forcing terms.

For the servo engineer, it is natural to visualize equations 1 by means of a signal-flow graph[26,27] (generalized block diagram), as shown in Fig. 1(A), where $1/s$ symbolizes integration. The signal-flow graph may be regarded as an abstract picture of the interconnections in an analogue computer[28] set up to simulate solutions of equations 1. Each little circle (node) represents one of the variables in equations 1. The r_i's are the independent variables and correspond to function generators in the analogue computer. The $\dot{c}_i$'s and c_i's are dependent variables and correspond to the output of adders and integrators respectively in the analogue computer. Branches directed toward a circle mean that the variable above the circle is a linear combination of those variables where the branches originate, each of the latter variables being weighted by the constant (or the operator $1/s$) which appears along the branch. Thus in Fig. 1(A)

$$\dot{c}_3 = 0 . c_1 + a_{32}c_2 + a_{33}c_3 + 1 . r_3$$

Note that it would be superfluous and even confusing to have branches connecting the $\dot{c}_i$'s among themselves.

EIGENVALUES AND EIGENVECTORS

It is well known that particular solutions of equations 1 have the form $c_i = $ constant $\epsilon^{s_\mu t}$. What are the permissible values of s? Substituting in the differential equation shows

$$s\mathbf{c} = A\mathbf{c} \tag{2}$$

The n homogeneous linear algebraic equations 2 have a nonzero solution if and only if their determinant vanishes

$$|A - sI| = 0 \tag{3}$$

where I is the unit matrix.

Expanding the determinant leads to an nth-degree algebraic equation in s. The n (not necessarily distinct) roots s_i of this equation are called the eigenvalues of the matrix A. Equation 3 is called the eigenvalue equation (or characteristic equation, or secular equation). Alternately, the s_i will be called the natural frequencies of the dynamic system represented

Fig. 1. Signal-flow graph representation of linear systems

A—A particular third-order system
B—Third-order system with real, distinct eigenvalues shown in normal co-ordinate phase space

LINEAR TRANSFORMATIONS

What happens to the eigenvalues under a linear transformation (change of variables, change of co-ordinate system)? A linear transformation may be written as

$$c = P_w w$$

or

$$w = P_w^{-1} c \tag{4}$$

where P_w is an $n \times n$ matrix with real elements and P_w^{-1} its inverse, $P_w P_w^{-1} = I$. A simple calculation shows

$$P_w^{-1} \frac{dc}{dt} = \frac{d(P_w^{-1}c)}{dt} = \frac{dw}{dt} = P_w^{-1} A c$$

$$= P_w^{-1} A P_w P_w^{-1} c = P_w^{-1} A P_w w$$

$$= B_w w \tag{5}$$

that the matrix A transforms as $B_w = P_w^{-1} A P_w$. Then

$$|B_w - sI| = |P_w^{-1} A P_w - sI|$$
$$= |P_w^{-1}(A - sI)P_w| = |P_w^{-1}| \cdot |A - sI| \cdot |P_w|$$
$$= |A - sI| = 0 \tag{6}$$

Thus the eigenvalues of A are the same as the eigenvalues of B_w, and A and B_w are therefore called similar matrixes. Two dynamic systems whose matrixes are A and B_w differ only superficially; they represent the same equations of motion, referred to different co-ordinate systems. It may also be said that A and B_w describe the motion of the system in two different phase spaces $\mathfrak{C}$ and $\mathfrak{W}$, in each of which a point is described by its orthogonal Cartesian co-ordinates. Of course, if a point in $\mathfrak{C}$ is given, the corresponding point in $\mathfrak{W}$ may be calculated by means of equations 4, and vice versa. The latter point of view will be used throughout the paper.

A particularly simple phase space is one whose matrix B_u contains only diagonal elements; it is the normal co-ordinate phase space $\mathfrak{U}$. The corresponding linear transformation P_u is called the normal co-ordinate or diagonalizing transformation. The signal-flow graph corresponding to $\mathfrak{U}$ has n disjoint parts, as shown in Fig. 1(B). The diagonal elements of B_u are the eigenvalues, and the differential equations in $\mathfrak{U}$ have the simple form (off-diagonal elements of B_u being zero)

$$u_i(t) = u_i(t_0) \epsilon^{s_i t}, \quad t \geq t_0 \ (i = 1, 2, \ldots n) \tag{7}$$

The initial conditions and the forcing function transform according to

$$u(t_0) = P_u^{-1} c(t_0)$$

and

$$r'(t) = P_u^{-1} r(t) \tag{8}$$

In other words, a separate differential equation may be written along each of the co-ordinate axes in $\mathfrak{U}$, with initial conditions and forcing function determined by equations 8. This means that in any m-dimensional subspace of $\mathfrak{U}$, formed by picking out $m \leq n$ co-ordinate axes, the trajectories are governed by an m-dimensional linear differential equation. Thus it is possible to isolate convenient sets of eigenvalues and define thereby smaller phase spaces which are contained in $\mathfrak{U}$.

In particular, assume that the transient behavior of the linear system may be approximated after some time $t_1 > t_0$ by two dominant eigenvalues. Then all the trajectories in $\mathfrak{U}$ may be approximated by the trajectories in that plane of $\mathfrak{U}$ which
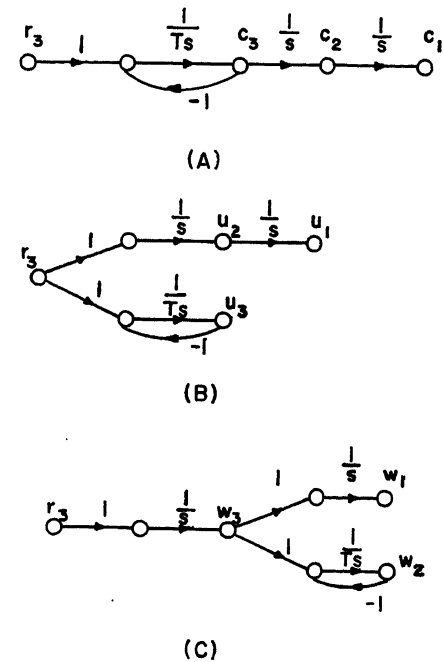
Fig. 2. Example of computation of transformation matrix

A—Original system
B—System in normal co-ordinate phase space
C—System in another phase space

contains the two eigenvalues. The trajectories in this plane, being governed by a second-order linear differential equation, are easily calculated. Finally, by projecting on some convenient plane in $\mathfrak{C}$, by means of P_u, an approximate but physically meaningful 2-dimensional picture of the transient behavior is obtained. This procedure represents one of the contributions of the present paper.

Note that if the eigenvalues are all real, then the eigenvectors in $\mathfrak{C}$ are simply the co-ordinate axes of $\mathfrak{U}$ transferred into $\mathfrak{C}$. If the dominant eigenvalues are real, then the trajectories in a suitable plane of $\mathfrak{C}$ are determined uniquely by the eigenvalues and the projections of the corresponding eigenvectors on the plane; computation of P_u, etc., is unnecessary.

Only very little use has been made so far of normal-co-ordinate transformations in studying nonlinear systems in the phase space.[12,30,31] This is partly because of the difficulty of obtaining the transformations by standard computational methods. A new point of view is necessary. This is developed next.

DIAGONALIZATION PROCESS RELATED TO TRANSFER FUNCTIONS

Complete diagonalization of the matrix A may not be possible or desirable. The process fails if A has multiple eigenvalues. Also, if any eigenvalues of A are complex,

by A^{29}; this is a plausible generalization of the fact that if the s_i are pure imaginary, the terms $\epsilon^{s_i t}$ give rise to sinusoids with angular frequency $|s_i|$. In the past, the concept of the eigenvalue has frequently been referred to also by the names characteristic value or root, latent root, proper value, etc.

Substituting the eigenvalues s_i given by equation 3 into equations 2, it is seen that any set of $n-1$ components of c may be expressed in terms of the remaining nth component of c. If s_i is real, the resulting expressions are simply the equations of a line through the origin in the phase space. This line is called the eigenvector of the matrix A. The names characteristic or proper vectors, principal axes or directions, eigenmodes, or principal modes have also been used in the past. Note that the eigenvector is a trajectory since equations 2 satisfy the original differential equations 1.

Fig. 3. Definition of saturation

then the diagonalizing transformation P_u may contain complex elements. For reasons of simplicity, it is best to work exclusively with real numbers. Complex numbers may be avoided if 2×2 submatrixes (with complex eigenvalues but real elements) are permitted to remain along the diagonal. The same device also allows consideration of double eigenvalues; eigenvalues of multiplicity three or higher are excluded in this paper.

The foregoing requirements are readily met and the computational steps for obtaining the matrix P_u are very much simplified by the following considerations. Notice that the diagonalizing transformation decomposes a complicated system into its natural frequencies (cf. equation 7). This is analogous to expanding a transfer function into partial fractions. The expansion can always be carried out if the natural frequencies are known; to compute P_u by any method, the natural frequencies must always be known also. As before, the process is considerably clarified if it is related to signal-flow graphs.

The following example illustrates the steps required in determining the components of P_u. Consider the transfer function $G(s)=1/s^2(Ts+1)$. The transfer function is represented by the signal-flow graph of Fig. 2(A), which defines the variables in $\mathcal{C}$. Since $G(s)$ has a double eigenvalue at $s=0$, complete diagonalization is not possible and only the eigenvalue $s=-1/T$ may be separated. This leads to Fig. 2(B).

A check on whether or not Fig. 2(B) was set up correctly is the verification of the known theorem

$$\text{trace } A \equiv \sum_i a_{ii} = \text{trace } B_u = \sum_i b_{ii}$$
$$= \sum \text{ eigenvalues}$$

Expanding the transfer functions relating c_1, c_2, and c_3 to r_3 leads to

$$\frac{C_1(s)}{R_3(s)} = \frac{1}{s^2(Ts+1)} = \frac{1}{s^2} - \frac{T}{s} + \frac{T^2}{Ts+1}$$

$$\frac{C_2(s)}{R_3(s)} = \frac{1}{s(Ts+1)} = \frac{1}{s} - \frac{T}{Ts+1} \qquad (9)$$

$$\frac{C_3(s)}{R_3(s)} = \frac{1}{Ts+1} = \frac{1}{Ts+1}$$

The coefficients in equations 9 make up the transformation matrix

$$c = P_u u = \begin{Vmatrix} 1 & -T & T^2 \\ 0 & 1 & -T \\ 0 & 0 & 1 \end{Vmatrix} u \qquad (10)$$

The foregoing result may be checked by computing the identity $AP_u = P_u B_u$, where A and B_u are the system matrixes in Figs. 2(A) and (B)

$$A = \begin{Vmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1/T \end{Vmatrix}$$

$$B_u = \begin{Vmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1/T \end{Vmatrix}$$

It is useful to have also the inverse P_u^{-1}, which is readily obtained by rearranging equations, since P_u is triangular

$$u = P_u^{-1} c = \begin{Vmatrix} 1 & T & 0 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{Vmatrix} c \qquad (11)$$

In general, an analytic form of P_u^{-1} can always be obtained directly by reversing the process of the partial fraction expansions shown in equations 9, which leads to equating of coefficients of powers of s. If P_u is given numerically and is of moderate order (up to about the fifth or sixth), a recent method advanced by Andree[32] is recommended for the computation of P_u^{-1} as particularly quick and simple.

A short calculation reassures that

$$P_u^{-1} A P_u = B_u$$

In setting up the signal-flow graphs in the phase spaces $\mathcal{C}$ and $\mathcal{U}$, care must be taken to select the variables c_i and u_i so that all transfer functions approach 0 as $s \to \infty$. Otherwise, the partial-fraction expansions will contain impulsive terms which would complicate the algebra.

This procedure for calculating the transformation matrix is obviously applicable not only to the signal-flow graph of Fig. 2(B) but to any other graph which contains the same natural frequencies, e.g., Fig. 2(C). The corresponding transformation matrix is

$$P_w = \begin{Vmatrix} 1 & -T & 0 \\ 0 & 1 & 0 \\ 0 & -1/T & 1/T \end{Vmatrix}$$

The following is a semirigorous justification of the transformation process described. First observe that a step input is equivalent, as far as the computation of trajectories is concerned, to a transient generated by some initial condition $c_0(t_0)$, provided no degeneracy is involved, i.e., a step input excites all of the natural frequencies of the system. Since the transformation $c = P_u u$ decomposes one particular forced solution $c_0(t)$ (not identi-

cally zero) in terms of its natural frequencies, it will decompose all others because of linearity. The proof of the last assertion follows easily by noting that solutions of equations 1 form a vector space.[22]

IDEALIZING ASSUMPTIONS

Definition of Saturation

Saturation is represented by an idealized curve; see Fig. 3. The saturation levels are normalized at ±1; the gain in the linear region is K_1. The letter f in the signal-flow graphs symbolizes the functional dependence depicted by Fig. 3.

Phase-Space Regions

In view of the foregoing definition, and in accordance with the author's generalized method of phase-plane analysis,[13] it is logical to subdivide the phase space into two distinct types of regions:

1. Saturation Regions. The trajectories are governed by the differential equations of the output element, excited by ±1. The poles and zeros of the transfer function of the output element are assumed to be known. The regions are labeled by $\pm$ signs according to the saturation polarity; their boundaries are shown by dash-dotted lines.

2. Linear Region. The trajectories are governed by the linear closed-loop differential equations; it is in each case necessary to determine approximately the closed-loop eigenvalues. The linear region is labeled L; its boundaries are shown by dash-dotted lines.

Permissible Inputs

The simplest and presently only workable method of introducing forcing functions in phase-plane analysis is by considering them as initial conditions.[13] In general, combinations of step and ramp inputs $r+\dot{r}t$ ($t\geq0$) are considered, although frequently the ramp component is set identically zero to simplify the discussion.

To identify step or ramp inputs with initial conditions, the following procedure is used: Investigate the dependence of critical points in $\mathcal{C}$ on r and $\dot{r}$, and then translate the co-ordinate system so as to leave the critical point always at the origin, if possible. For instance, if for a step input $c_1 = r$ in the steady state, the variable c_1 is changed to $r-c_1=e$; the steady-state equilibrium point is then always at $e=0$. This process is possible only if the error in $\mathcal{C}$ remains finite, which is not the case, e.g., when a nonzero position-error servomechanism is subjected to a ramp input $\dot{r}t$. The method pre-

(A)

(B)
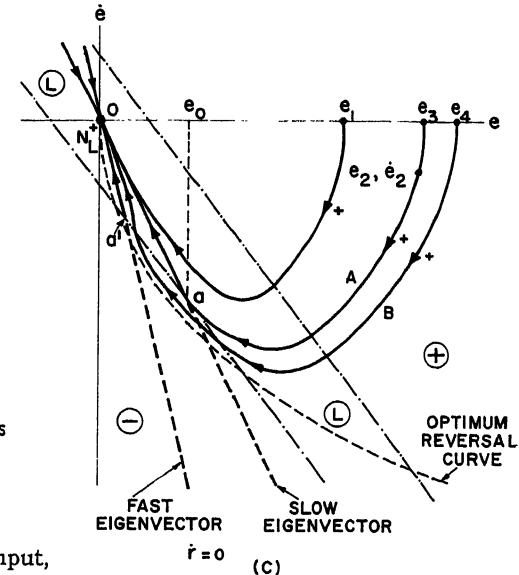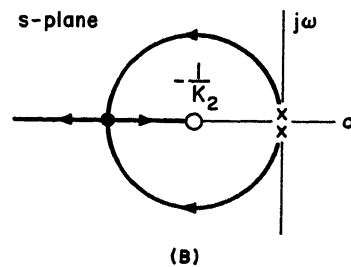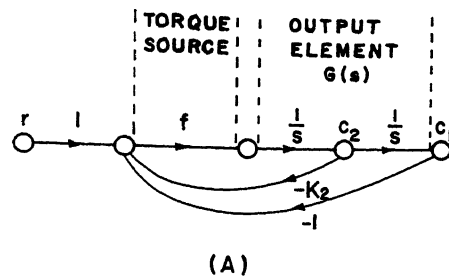
Fig. 4. Example 1

A—Signal-flow graph
B—Root locus
C—Phase-plane trajectories

(C)

---

sented for the analysis of high-order systems shows excellent promise also for the treatment of arbitrary (but nonrandom) inputs.

Analysis of Second-Order Saturating Systems

EXAMPLE 1. POSITION CONTROL SERVO CONSISTING OF A SATURATING TORQUE SOURCE DRIVING PURE INERTIA

Consider an idealized system with the following parts: a torque source (motor, etc.) whose dynamics are negligible but which is subject to saturation with a transfer function f, and a purely inertial linear output element with a transfer function $G(s) = 1/s^2$. To assure stability, it is necessary to introduce some form of compensation, for instance, a lead network or velocity (tachometer) feedback. Any physical lead network possesses not only a zero but also a pole which makes the system third-order; discussion of the effect of the pole is postponed until Example 4. If velocity feedback is chosen, the signal-flow graph for the system is as shown in Fig. 4(A).

In the saturation region, the feedback is not effective; the system is simply a double integration excited by ± 1. The trajectories are parabolas. The optimum reversal curve for both steps and ramps is $2e = -\dot{e}|\dot{e}|$; see Appendix II.

In the linear region, the system is governed by the following differential equation, written down by inspection from the graph

$$\frac{dc}{dt} = \left\| \begin{array}{cc} 0 & 1 \\ -K_1 & -K_1K_2 \end{array} \right\| c + \left\| \begin{array}{c} 0 \\ K_1r \end{array} \right\| = Ac +$$

constant

The eigenvectors of A are the straight lines

$$s_ic_1 = -\frac{K_1}{K_1K_2 + s_i} c_1 = c_2 \quad (i = 1, 2) \quad (12)$$

The locus of the eigenvalues s_i, i.e., the locus of the closed-loop poles, is shown in Fig. 4(B). Increasing the gain K_1 results first in two complex and then in two real roots; one of the real roots goes to the zero and the other to $-\infty$ as $K_1 \to \infty$.

The boundaries of the linear region are the straight lines

$$-K_2c_2 + (r - c_1) = \pm 1/K_1$$

To eliminate dependence on the input, set $r - c_1 = e$ and $\dot{r} - c_2 = \dot{e}$; then

$$e + K_2\dot{e} = \pm 1/K_1 + \dot{r}K_2 \quad (13)$$

With this change of variables, the critical point in the linear region will remain at the origin for all step and ramp inputs.

With the knowledge of the optimum reversal curve, the eigenvectors, the root locus, and the eigenvalues and boundaries of the linear region, a complete solution of the problem is possible. Restrict the input to steps ($\dot{r}_2 \equiv 0$), and consider first the case of real eigenvalues. Noting that the slope of the slow eigenvector, defined in Appendix I, is always larger than the slope of the linear region, i.e.,

$$-s_1 > 1/K_2 \quad (s_1 > s_2)$$

it is clear that, for sufficiently large K_2, the situation in the $(e, \dot{e})$ phase plane must be as depicted in Fig. 4(C). A transient trajectory will remain in the linear region and go to zero without overshoot if it moves sufficiently close to the slow eigenvector after entering the linear region. The largest such trajectory is denoted by A in Fig. 4(C). Trajectory A may have resulted from a step error e_3 or a combination of step and ramp errors e_2 and $\dot{e}_2$. If the error is larger (e_4), the resulting trajectory B may still reach the origin without overshoot if it re-enters the linear region in the segment between the two eigenvectors (a and a' in the figure), in which case it will become asymptotic to the slow eigenvector from the other side.

If the magnitude of the error is further increased, the response will deteriorate as the trajectories enter and leave the linear region more and more times. This is because the effective reversal of saturation takes place later and later than required by the optimum reversal curve. The question arises as to how such poor performance can be minimized by proper selection of K_2.

Fig. 4(C) shows that the largest error for which the trajectory still remains in the linear region is roughly proportional to e_0, the intersection of the slow eigenvector with the boundary of the linear region. From the geometry and equations 12 and 13

$$-e_0 = 1/K_1(1 + K_2s_1) = 1/K_1 \times$$

$$\left[1 - K_2\left(\frac{K_1K_2}{2} - \sqrt{\frac{K_1^2K_2^2}{4} - K_1} \right) \right]$$

$$= 1/K_1 \left[1 - \frac{K_1K_2^2}{2}\left(\frac{2}{K_1K_2^2} + \right. \right.$$

$$\left. \left. \frac{2}{K_1^2K_2^4} + \frac{4}{K_1^3K_2^6} + \cdots \right) \right]$$

Hence if $4/K_2^2 \leq K_1 < \infty$

$$e_0 \approx K_2^2 \quad (14)$$

Equation 14 reveals the dilemma of the situation: to have a good step response to large errors, e_0 should be made as large as possible; on the other hand, since $-s_1 \approx 1/K_2$, this would result in very slow response of the system after it leaves saturation, or for small inputs. The designer must compromise between the requirements of nonoscillatory response to large and sufficiently fast response to small errors. For a previous treatment of the same problem, see Kendall and Marquardt.[11]

When ramp inputs $\dot{r}_2 \neq 0$ are also admitted in addition to steps, the location of the linear region in the $(e, \dot{e})$ phase plane will shift. This represents a further departure from the optimum reversal curve requirements. The shift can be eliminated only by measuring the derivative of the input, i.e., using a lead network which operates on the error signal. The ensuing analysis is similar to the foregoing.

The discussion becomes somewhat less straightforward if the eigenvalues in the

---

(A)

(B)
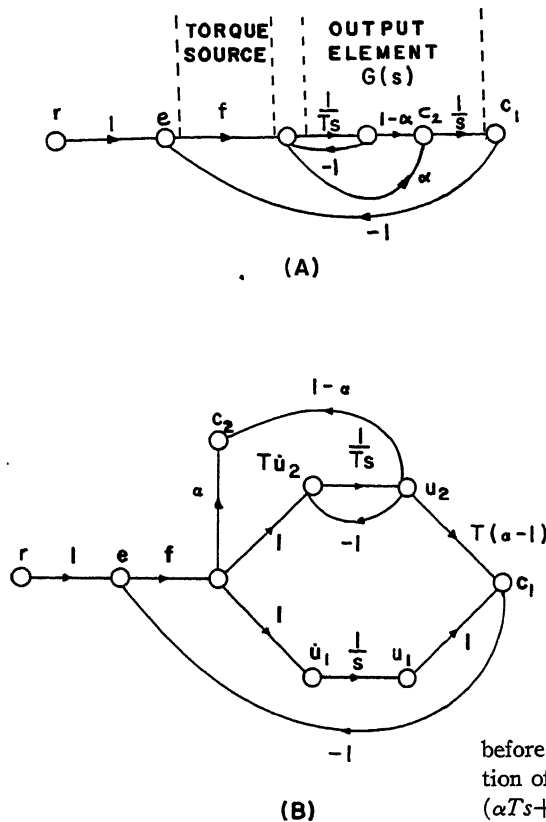
Fig. 5. Example 2

A—Signal-flow graph in c
B—Signal-flow graph in $\mathfrak{U}$

linear region are complex. Then the simplest representation of the trajectories is a logarithmic spiral, see Appendix I, which has to be transformed from the normal co-ordinate system to the actual phase plane $(e, \dot{e})$. In the present example, however, the root locus shows that the time of response will decrease with the gain until about critical damping is reached. Hence it is not desirable in this case to keep K_1 so small that complex eigenvalues result.

The most important aspects of nonlinear system design are now apparent. It is meaningless to attempt to formulate hard and fast criteria for the adjustment of the parameters K_1 and K_2, such as the M_p criterion, phase margin, etc., used in linear systems, because the whole design must be predicated on some compromise between the desired responses to particular classes of signals. At the present state of the theory, only step or ramp-type signals can be treated and this restricts the analysis to consideration of signal magnitudes only. Moreover, linear compensation may be inadequate altogether and a better approximation to the optimum reversal line may be necessary.

Example 2. Servo with Zero in Open-Loop Transfer Function

Consider the same type of system as

before, but assume that the transfer function of the output element is now $G(s) = (\alpha Ts+1)/s(Ts+1)$; see Fig. 5(A). It will be convenient to carry out the entire analysis in terms of the normal co-ordinates of the saturation region. The appropriate transformation is obtained by expanding into partial fractions the transfer functions

$$G(s) = \frac{C_1(s)}{\pm 1/s} = \frac{1}{s} + \frac{T(\alpha-1)}{Ts+1}$$

and

$$sG(s) = \frac{C_2(s)}{\pm 1/s} = \alpha + \frac{1-\alpha}{Ts+1}$$

Fig. 5(B) is the revised signal-flow graph showing the old variables c_1 and c_2 in terms of the new variables u_1 and u_2.

The dependence of the critical points in the linear region of the (u_1, u_2) phase plane on the inputs may be deduced from steady-state considerations as follows. For steps, $r-c_1=e=0$ implies $r-u_1=0$, $u_2=0$. For combinations of steps and ramps, $r-c_2=\dot{e}=0$ together with $e=\dot{r}/K_1$ implies $r-u_1=\dot{r}[T(\alpha-1)+1/K_1]$ and $\dot{r}-u_2=0$. Hence it is logical to work in the $(r-u_1, \dot{r}-u_2)$ phase plane which is similar to the usual $(e,\dot{e})$ phase plane. For simplicity, the following discussion is limited to step inputs, $\dot{r}\equiv0$, since then the critical point is always at the origin.

The trajectories in the saturation region are given by $u_2=\pm[1\mp$ constant $\exp(\mp u_1/T)]$ and have the same qualitative shape as in case 2, Appendix II.

The boundaries of the linear region are the straight lines

$$(r-u_1) - T(\alpha-1)(-u_2) = \pm 1/K_1 \qquad (15)$$
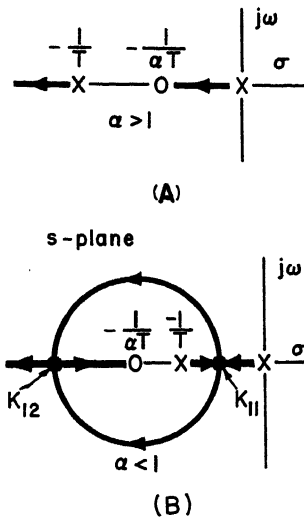
Two distinct types of root loci are possible, depending on whether $\alpha>1$ or $\alpha<1$, and are shown in Figs. 6(A) and (B). As in the previous example, only real roots will be of interest, and thus the critical point at the origin will always be a stable node $N_L{}^+$. Assuming first $\alpha>1$, consider the effect of gain adjustments on the transient behavior in the linear region. The equation of the eigenvectors is

$$(1+s_i/K_1)(r-u_1) = -T(\alpha-1)(-u_2)$$
$$(i=1, 2) \quad (16)$$

As the eigenvalues vary in accordance with the root locus, the eigenvectors rotate around the origin. The slow eigenvector starts on the $(r-u_1)$ axis, swings into the second and fourth quadrants, and approaches the line $(r-u_1)=-T(\alpha-1)(-u_2)$ as $K_1\to\infty$. The fast eigenvector starts on the $(-u_2)$ axis, swings into the first and third quadrants, and approaches the line $(r-u_1)=T(-u_2)$. See Fig. 7(A). It is also necessary to know whether the intersection A of the slow eigenvector with the line $-u_2=\pm1$ falls inside or outside the linear region. Using equations 15 and 16, the condition $a<b$ becomes

$$T(\alpha-1)+1/K_1<-1/s_1, (s_2<s_1) \qquad (17)$$

Since from the eigenvalue equation $1/K_1+\alpha T=T_1+T_2$, where $s_1=-1/T_1$, and $s_2=-1/T_2$ are the closed-loop poles in Fig. 6(A), inequality 17 becomes

$$T-T_2>0$$

In view of the root locus, the inequality is satisfied for all K_1. It is now possible to sketch a typical trajectory by inspection, as had been done in Fig. 7(A), without the danger of having omitted any important details.

If $\alpha<1$, the only essential difference from the previous case is because of the new root locus, Fig. 6(B). By carrying out simple calculations, analogous to the foregoing, the phase-plane sketches shown in Figs. 7(B) and (C) are arrived at. As $K_1\to\infty$, the eigenvectors first swing toward one another in the second and fourth quadrants as in Fig. 7(B) and coincide when $K_1=K_{11}$. If K_1 is further increased, the eigenvalues become complex conjugate, which case is not analyzed. When $K=K_{12}$, the eigenvectors again coincide, this time in the first and third quadrants and separate with K_1 increasing still further, as shown in Fig. 7(C).

Note that in all the computations which determine the qualitative features of Fig. 7 and, in particular, the location of the point A and the position of the eigenvectors, it is only necessary to use the eigenvalue equation and the qualitative form of the root locus No explicit com-

Fig. 6. Example 2; root loci

A—$\alpha>1$
B—$\alpha<1$

Kalman—Second and Higher Order Saturating Servomechanisms

(A)

$(-u_2)$

N_L^+

a b

$(r-u_1)$

$\ominus$ $\oplus$

-1

A

FAST EIGENVECTOR

$\ominus$ L

$e<0$ $e>0$ SLOW EIGENVECTOR

$\dot{r}=0$ $\alpha>1$

(B)

$(-u_2)$

$\ominus$ L

N_L^+

O

$(r-u_1)$

-1 $\oplus$

$+$

A

$e<0$ $e>0$ SLOW EIGENVECTOR

FAST EIGENVECTOR

$\dot{r}=0,\ 0<K_1<K_{11},\ \alpha<1$

(C)

$(-u_2)$ L

$\ominus$

O

FAST EIGENVECTOR N_L^+

$(r-u_1)$

A $\oplus$ $+$

-1

SLOW EIGENVECTOR

$e<0$ $e>0$

$\dot{r}=0,\ K_{12}<K_1<\infty,\ \alpha<1$

Fig. 7. Example 2, phase-plane trajectories

A—$\alpha>1$
B—$\alpha<1,\ 0<K_1<K_{11}$
C—$\alpha<1,\ K_{12}<K_1<\infty$
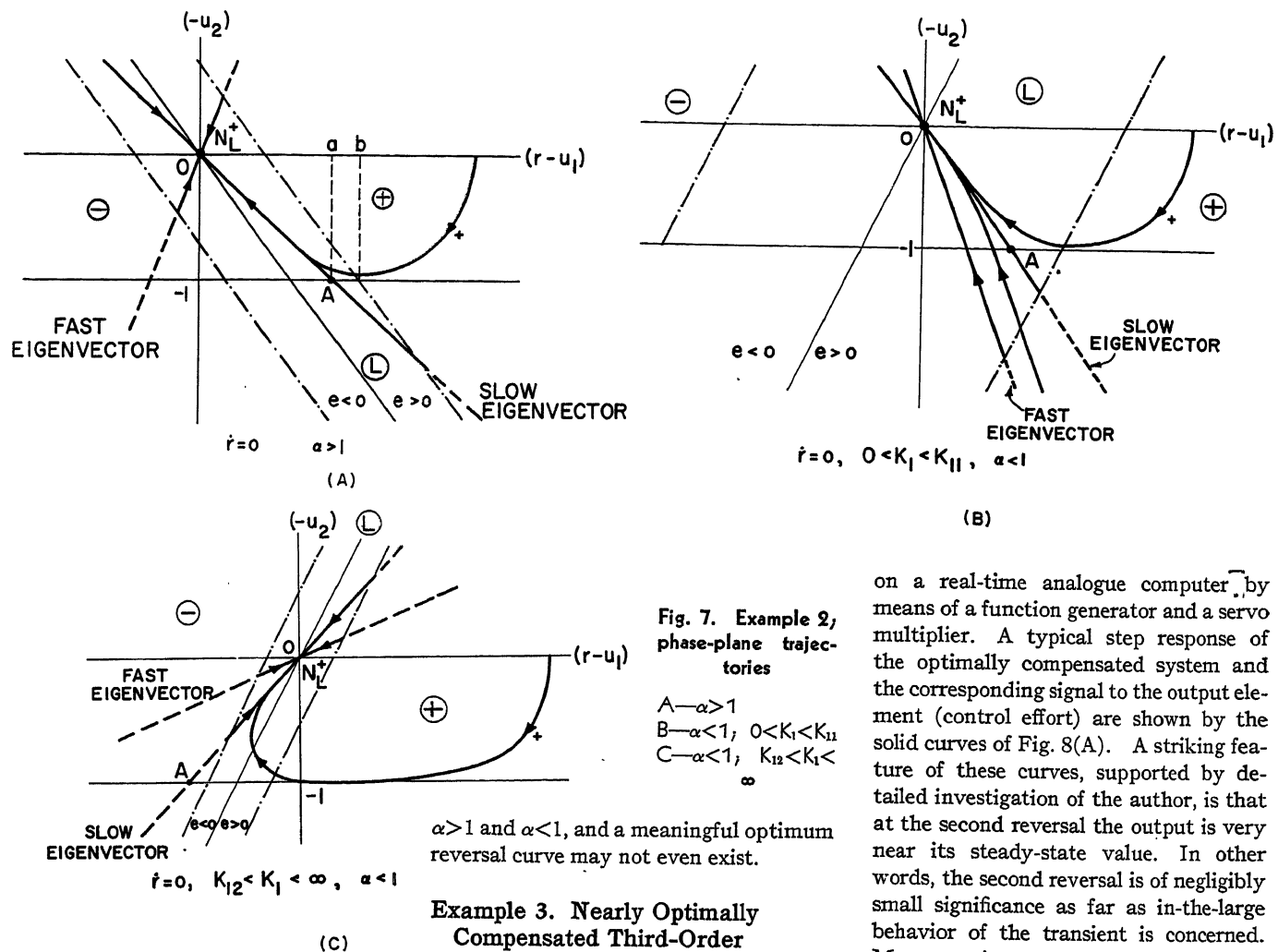
$\alpha>1$ and $\alpha<1$, and a meaningful optimum reversal curve may not even exist.

Example 3. Nearly Optimally Compensated Third-Order Saturating Servomechanism

Analogue Computer Study

The following is a summary of an analogue computer study of the optimum control of a saturating third-order servomechanism, carried out by J. L. Preston.[14] The problem is characterized by a transfer function $G(s) = 1/s^2(Ts+1)$. Such a problem may arise when it is impossible to reverse instantaneously the torque in the system of Example 1 because of the unavoidable energy storage in the torque source.

As pointed out in Appendix II, the optimum nonlinear compensation scheme calls for three reversals of the signal to the output element: the first occurs when the trajectory intersects a reversal surface, the second when the trajectory moving in the reversal surface intersects a reversal line, and the third on reaching the origin of the $(e, \dot{e}, \ddot{e})$ phase space. The equations of the reversal surface and line may be derived by utilizing a normal coordinate transformation[12] or by computing the solutions of the output-element differential equation backwards in time, using the substitution $t' = -t$.

The reversal surface was approximated

putation of either the eigenvalues or the corresponding value of K_1 is necessary.

The end results of the analysis shown in Fig. 7 contain few surprises. The optimum relay servomechanism theory is not applicable in this case because of the presence of the zero in the transfer function; see Appendix II. The large-signal response in Fig. 7(A) is acceptable and quite insensitive to wide variations in K_1 or α. The response in Fig. 7(B) is very poor; it may be improved by increasing K_1 which leads to Fig. 7(C). The transient in this figure will always overshoot, but the effect is easily minimized by choosing a large value of K_1.

As a final point, notice the intrinsic difference between putting the term $\alpha Ts+1$ before or after the saturation. In the first case, the servo has error-rate feedback and the linear region may be adjusted so that it will contain the entire optimum reversal curve between the limits $\dot{e}=\pm 1$. As shown by Kendall and Marquardt[11] this is equivalent to setting $\alpha\approx 1-ln\ 2=0.307$. In the second case, discussed in the foregoing, the response may be entirely acceptable for both

on a real-time analogue computer by means of a function generator and a servo multiplier. A typical step response of the optimally compensated system and the corresponding signal to the output element (control effort) are shown by the solid curves of Fig. 8(A). A striking feature of these curves, supported by detailed investigation of the author, is that at the second reversal the output is very near its steady-state value. In other words, the second reversal is of negligibly small significance as far as in-the-large behavior of the transient is concerned. Moreover, in any practical system it is necessary to provide a small linear region near zero error since otherwise small inaccuracies in the reversal surface and/or noise would seriously impair the steady-state performance. The second reversal will then be brought about while the system behaves linearly.

In view of these considerations, a simplified compensation was arrived at. It consists of: 1. a small linear zone, of width $2/K_1=2T$; and 2. a reversal curve in the $(e, \dot{e})$ phase plane, which is simply the intersection of the reversal surface with $\ddot{e}=-1$ in the fourth quadrant and $\ddot{e}=+1$ in the second quadrant.

Step 2 is based on the assumption that $\ddot{e}$ has reached its steady-state value at the time of the first reversal; this is certainly an excellent approximation for large signals and is of no interest for small signals since then the trajectory stays in the linear region. The dash-dotted curves in Fig. 8(A) show the response obtainable with this nearly optimum compensation. The response follows the optimum transient curve until entering the linear region, when a slight oscillation (somewhat exaggerated in the figure) develops. The last part of the transient is approximately the same for all magnitudes of the input.

The reversal curve in the $(e, \dot{e})$ phase plane is shown by the solid curve in Fig. 8(B). It is seen that the signal to the third-order output element must be reversed much earlier than in the case of the second-order transfer function $1/s^2$ (dash-dotted line), since the actual reversal of the acceleration $-\ddot{e}=c_3$ takes place with a lag $1/(Ts+1)$. The signal-flow graph of the nearly optimum system is shown in Fig. 8(C) where the equation $g(e)+K_2\dot{e}=0$ represents the reversal curve.

Finally, the dashed curves in Fig. 8(A) show the response of a saturating linear system, with the parameters adjusted experimentally so as to minimize the performance-criterion expression[33]

$$\int_0^\infty t\,|e|\,dt$$

in response to a step during entirely linear operation. The following optimum values were found: loop gain $K_1=0.186/T^2$ and velocity-feedback gain $K_2=3.76T$. Notice that while the nearly optimum control effort, dash-dotted curve, is quite different from the optimum control effort, solid curve, the output responses are very nearly the same. By contrast, although the dash-dotted control effort closely resembles the dashed one, the corresponding output responses are very different as the linear saturating system has large overshoot and effectively twice as long speed of response, because the acceleration is reversed too late.

SIMPLIFIED ANALYSIS AND DESIGN
PROCEDURE

The preceding discussion shows that very nearly optimum response may be achieved by using an appropriate reversal curve for the first reversal, and satisfying the need for a second reversal near the origin by leading the trajectory into a small linear region. This suggests the following simplified derivation of the reversal curve which at the same time lends excellent insight into the principal quantitative questions involved. It is simply the implementation of the approximation stated under point 3 in the section entitled "Principles of Analysis."

First, calculate the normal co-ordinate transformation for the transfer function $G(s)=1/s^2(Ts+1)$. This was discussed in the foregoing in connection with Figs. 2(A) and (B). The transformation matrix P_u is given by equation 10.

In $\mathfrak{U}$, the plane (u_1, u_2) containing the transfer function $1/s^2$ is optimally compensated in the usual fashion, while along the u_3 axis the transient is assumed to have reached its steady-state value at the time of the first reversal. The assumption
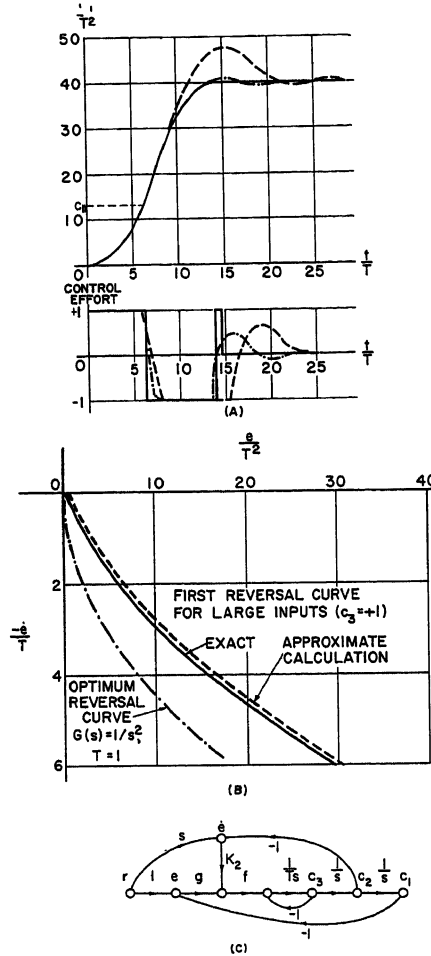
Fig. 8. Example 3

A—Transient responses
B—Reversal curves
C—Signal-flow graph of nearly optimally compensated system

concerning u_3 was also implicit in deriving the first reversal curve from the reversal surface in the analogue computer study. Optimum compensation in the (u_1, u_2) plane is not identical with optimum compensation in the $(e, \dot{e}, \ddot{e})$ phase space; the error involved in the procedure, however, is only the neglection of the second reversal and, as mentioned in the foregoing, this is totally unimportant.

The optimum reversal curve in the (u_1, u_2) plane consists of two half-parabolas

$$2u_1 = -u_2|u_2| \tag{18}$$

Substituting the transformation equation 11 into equation 18 and $u_3=\pm1$ leads to two equations in c_1, c_2, and c_3 from which c_3 may be eliminated. Finally, a shift in the origin yields the analytic expression of the reversal curve in the $(e, \dot{e})$ phase plane

$$2e = -(|\dot{e}|^2+4T|\dot{e}|+T^2)|\dot{e}|/\dot{e} \tag{19}$$

Geometrically, the foregoing calculation is equivalent to projecting the re-

versal curve from $\mathfrak{U}$ onto the (c_1, c_2) plane of $\mathfrak{C}$. Equation 19 is shown by the dashed curve in Fig. 8(B); agreement with the solid reversal curve, which is the intersection of the exact reversal surface with $c_3=+1$, is very close.

Application of the normal co-ordinate transformation eliminates the necessity of computing the reversal surface. The first reversal curve is determined very quickly from the well-known second-order cases. It is not even necessary to express the reversal curve in the $(e, \dot{e})$ plane analytically; by means of P_u^{-1}, it may be transferred from the (u_1, u_2) plane point by point.

Other relevant information is equally readily calculated:

1. Time of response: This must be roughly the time of response of the optimum second-order system in the (u_1, u_2) plane, plus the time required to reduce u_3 from ±1 to zero on reversing the relay polarity, about T seconds. Hence the approximate time of response to step r is

$$2\sqrt{|r|}+T \tag{20}$$

2. Error at first reversal: In the (u_1, u_2) phase plane the reversal occurs when $(r-u_1)=r/2$, $u_2=-\sqrt{|r|}$. Hence

$$|e_{11}| = |r-c_{11}| = |r|/2+T\sqrt{|r|}+T^2$$

These results also agree closely with Fig. 8(A).

Examples of Third–Order Saturating
Control Systems

GENERAL

Example 3 showed that the crucial difference between the analysis of second- and high-order systems lies in determining how the trajectories enter and leave the two different saturation regions. When there is also a linear region, the following further considerations arise:

1. The transformation from $\mathfrak{C}$ to the saturation-region normal co-ordinate phase space $\mathfrak{U}$ is not the same as from $\mathfrak{C}$ to the linear-region normal co-ordinate phase space $\mathfrak{V}$. Hence, while all the approximations concerning the neglect of fast transient terms must be carried out in the normal co-ordinate phase spaces, to have a complete picture of the solutions the results of the approximations are to be projected onto some one phase plane in any of the spaces $\mathfrak{U}$, $\mathfrak{V}$, or $\mathfrak{C}$. It will be usually convenient to refer everything to the (u_1, u_2) or $(e, \dot{e})$ phase planes. Thus, the analysis leads to an approximately equivalent second-order nonlinear system, defined in any convenient phase plane.

2. The linear region in the equivalent phase plane will have two different types of boundaries:

a. Points where trajectories leave one of the saturation regions and enter the linear region, derived under the assumption

(A)

(B)
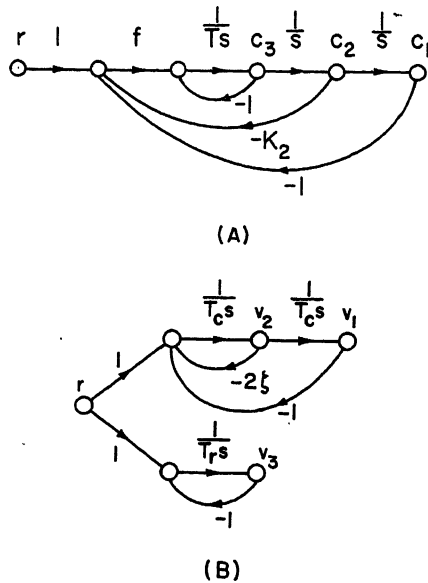
Fig. 9. Example 4

A—Signal-flow graph in $\mathcal{C}$
B—Signal-flow graph in $\mathcal{U}$

---

that the fast transients in $\mathcal{U}$ have reached their steady states.

b. Points where trajectories leave the linear region and enter one of the saturation regions, derived under the assumption that the fast transients in $\mathcal{U}$ have reached their steady states.

A geometric picture is helpful here. During saturation, all trajectories in $\mathcal{C}$ tend asymptotically to a certain plane which corresponds to, say, the (u_1, u_2) plane in $\mathcal{U}$. The intersection of this plane with that boundary plane of the linear region in $\mathcal{C}$ where the trajectories enter is a line. The projection of this line on the equivalent phase plane is the boundary of type a. During linear operation, the trajectories in $\mathcal{C}$ tend to a different plane which corresponds to, say, the (v_1, v_2) plane in $\mathcal{U}$. The intersection of this plane with that boundary plane of the linear region in $\mathcal{C}$ where the trajectories leave, projected on the equivalent phase plane, is the boundary of type b.

The following examples have been chosen to illustrate these considerations and also to reveal some of the peculiar differences which distinguish higher order systems from their second-order counterparts.

EXAMPLE 4. VELOCITY FEEDBACK IN THIRD-ORDER SERVO

This is the problem of the preceding section, but here the interest centers on the effect of linear velocity feedback. The signal-flow graph is shown in Fig. 9(A). The appropriate transformations from $\mathcal{C}$ to $\mathcal{U}$ and vice versa have already been derived in equations 10 and 11. The

---

boundaries of the linear region referred to $\mathcal{U}$ are

$$r - c_1 - K_2 c_2 = r - u_1 + T u_2 - T^2 u_3 - K_2(u_2 - T u_3) = \pm 1/K_1 \quad (21)$$

Assuming that u_3 is very near its steady-state value, i.e., substituting $u_3 \cong \pm 1$ in equation 21, gives the approximate boundary lines (type a) where the trajectories enter the linear region

$$(r - u_1) + (K_2 - T)(\dot{r} - u_2) = \mp T(K_2 - T) \pm 1/K_1 + (K_2 - T)\dot{r} \quad (22)$$

To find the boundaries where the trajectories leave the linear region (type b), it is necessary to have the transformation between $\mathcal{C}$ and $\mathcal{U}$. For convenience, the co-ordinates in $\mathcal{U}$ are selected as shown in Fig. 9(B). The quantities T_r, T_c, and ζ are to be determined from the root locus for any particular value of K_1; if all the eigenvalues are real, $-1/T_r$ will denote the most negative eigenvalue. The transformation matrixes are

$$c = P_v v = \frac{1}{T_r^2 + T_c^2 - 2\zeta T_r T_c} \times \begin{Vmatrix} T_c^2 & -2\zeta T_r T_c & -T_r T_c & T_r^2 \\ T_r & T_c & -T_r \\ -1 & T_r/T_c - 2\zeta & 1 \end{Vmatrix} v \quad (23)$$

and

$$v = P_v^{-1} c = \begin{Vmatrix} 1 & T_r & 0 \\ 0 & T_c & T_r T_c \\ 1 & 2\zeta T_c & T_c^2 \end{Vmatrix} c \quad (24)$$

From equation 23, the boundaries of the linear region referred to $\mathcal{U}$ are

$$r - c_1 - K_2 c_2 = \pm 1/K_1 = r - \frac{[T_c^2 + T_r(K_2 - 2\zeta T_c)]v_1 + T_c(K_2 - T_r)v_2 + T_r(T_r - K_2)v_3}{T_r^2 + T_c^2 - 2\zeta T_r T_c} \quad (25)$$

Using the linear transformation $v = P_v^{-1} P_u u$ and the assumption that $v_3 \cong 0$ when the trajectories leave the linear region, equation 25 may be referred to $\mathcal{U}$ and u_3 eliminated. Thus the equation of the boundaries of type b in the (u_1, u_2) phase plane is

$$(TK_2 - 2\zeta TT_c + T_c^2)(r - u_1) + T_c^2(K_2 - T)(\dot{r} - u_2) \pm (T^2 - 2\zeta TT_c + T_c^2)/K_1 = T_c^2(K_2 - T)\dot{r}$$

Using the eigenvalue equation, T_c and ζ may be eliminated by introducing T_r

$$(T_r^2 K_1 + 1)(r - u_1) + (K_2 - T)(\dot{r} - u_2) \pm [1/K_1 + (T + T_r - K_2)T_r] = (K_2 - T)\dot{r} \quad (26)$$

During saturation, the situation in the $(r - u_1, \dot{r} - u_2)$ phase plane is similar to that in the $(e, \dot{e})$ phase plane of example 1, but now K_2 is replaced by $(K_2 - T)$ in equation 13. If $1/K_1 \approx 0$, i.e., the linear

---

$T < K_2 < 9T$

(A)

$K_2 > 9T$

(B)
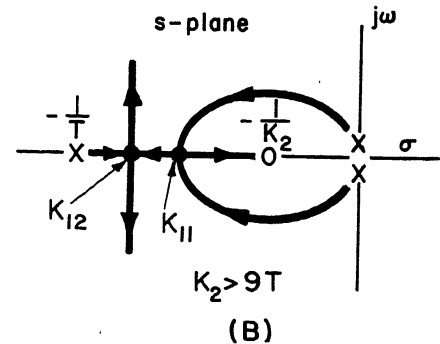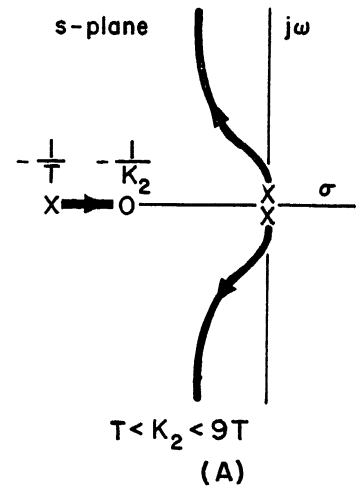
Fig. 10. Example 4; root loci

A—$T < K_2 < 9T$
B—$K_2 > 9T$

---

region is very small, the system becomes a relay servo. The condition for this relay servo to be stable for all inputs is a negatively sloping reversal line in the $(r - u_1, \dot{r} - u_2)$ plane

$$K_2 > T \quad (27)$$

Now examine the linear region. If $K_2/T > 1$, two qualitatively different root loci are possible depending on the magnitude of K_2/T, as shown in Figs. 10(A) and (B). In both cases, the eigenvalues remain in the left-half plane for all values of $K_1 > 0$. Hence inequality 27 which guarantees the stability in the saturation region also implies stability in the linear region. Conversely, if $K_2/T < 1$, i.e., the pole and zero in Fig. 10(A) are interchanged, which is equivalent (approximately) to reflecting about the $j\omega$ axis the root loci starting at the double pole at the origin, there will be two complex eigenvalues in the right-half plane for any $K_1 > 0$. Thus violation of inequality 27 implies instability in both the saturation and linear regions.

Since equation 26 was derived by the assumption $v_3 \cong 0$, it is valid only if the real eigenvalue $-1/T_r$ represents a transient which can be neglected in comparison with the other two poles. This is

EIGENVECTOR

$\dot{r}=0,$ $K_2>9T,$ $K_1=K_{11}$
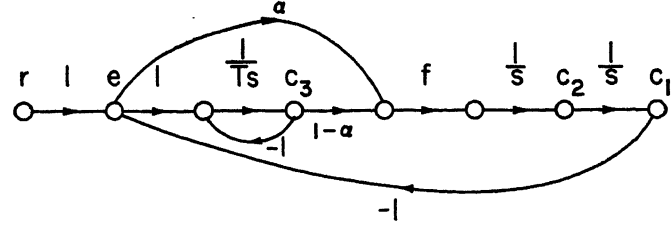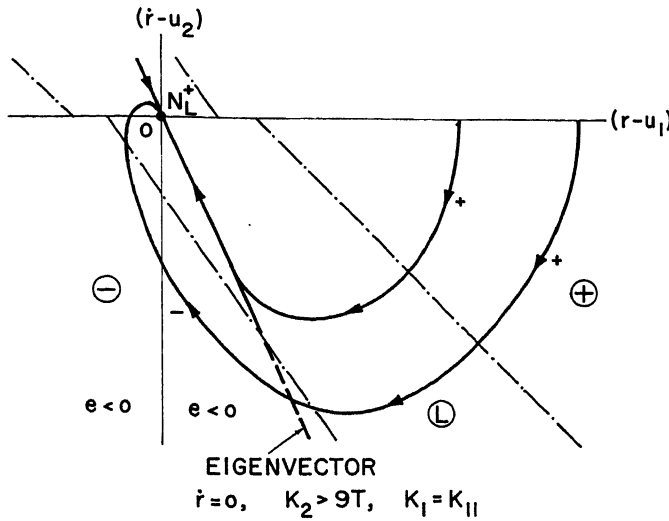
Fig. 11 (left). Example 4, phase-plane trajectories

Fig. 12 (above). Example 5, signal-flow graph in $\mathbb{C}$

true in Fig. 10(A) for $K_1>0$ and in Fig. 10(B) for $0<K_1<K_{12}$.

If, however, $K_1>K_{12}$ in Fig. 10(B), the obvious assumption would be $v_1=v_2\cong0$. If the assumption is correct, it means that the trajectories will have synchronized to the slow eigenvector before leaving the linear region. But if the trajectory has already synchronized to the eigenvector, it will move along the eigenvector closer and closer to the origin, and therefore cannot leave the linear region. It should be remembered that the boundary of the linear region here is a plane; if it were a more complicated surface, the conclusion would not necessarily hold. In reality, if the trajectories start sufficiently far from the origin, however, they will not synchronize to the eigenvector before leaving the linear region, so that the assumption $v_1=v_2\cong0$ is untenable and a more detailed analysis is necessary. The latter case is too complicated to be considered in this paper.

Fig. 11 shows the boundaries of the linear region in the $(r-u_1,\dot{r}-u_2)$ phase plane. As in example 3, the usual optimization theory for the transfer function $1/s^2$ applies in this plane.

To appraise the situation in the linear region, it is necessary to project the dominant eigenvector from $\mathbb{C}$ into the $(r-u_1, \dot{r}-u_2)$ plane. In $\mathbb{C}$ the eigenvectors are defined by the equations

$s_ic_1=c_2$

$s_ic_2=c_3$

Hence, making use of P_u

$s_i(u_1-Tu_2+T^2u_3)=u_2-Tu_3$

$s_i(u_2-Tu_3)=u_3$ (28)

Eliminating u_3 from equation 28 gives the projection of the eigenvector on the $(r-u_1, \dot{r}-u_2)$ plane

$s_i(r-u_1)=\dot{r}-u_2$ (29)

The fact that this equation is the same in both the $(r-u_1, \dot{r}-u_2)$ and $(e, \dot{e})$ phase planes is merely an accident.

It remains to investigate the quantitative effects of changes in K_1 and K_2/T. If the root locus of Fig. 10(B) applies, the complex roots which arise when $0<K_1<K_{11}$ are of no interest, for the same reason as in example 1. If $K_{11}\leq K_1<K_{12}$, the simplified analysis is admissible. Letting $K_1=K_{11}$ for convenience ($\zeta=1$) and bearing in mind that $0<T<K_2/9$, manipulation of the eigenvalue equation shows

$K_2/3<T_c<K_2/2$ and $1<1+T_r^2K_1<4/3$ (30)

Comparing equations 22, 26, 29, and 30 shows that the relative slopes of the boundaries of the linear region and of the dominant eigenvector must be as sketched in Fig. 11. Even large variations of K_2/T cannot affect the situation. The resulting transient response is not very good. Some improvement might be had by increasing K_1 beyond K_{12}, but too large a value of K_1 would result in the trajectories cutting across the linear region because of more and more poorly damped complex poles, and transients caused by large errors may overshoot several times. Thus it appears that too much velocity feedback is not beneficial.

If the root locus of Fig. 10(A) applies, the analysis may be carried analogously, although the spiral trajectories in the linear region corresponding to the complex eigenvalues are not as readily handled quantitatively. In this case again, if K_1 is increased indefinitely, more and more trajectories will be leaving the linear region; see Figs. 15(D), (E), and (F).

With a moderate value of K_1, fairly satisfactory large-signal performance can be expected. The precise optimum relation between K_1 and K_2/T is probably best obtained by an analogue computer

study guided by the foregoing considerations.

EXAMPLE 5. LEAD-NETWORK COMPENSATION

Now consider the lead-network compensation of the servomechanism of example 1. The development of the analysis is similar to but not identical with the preceding example. The lead network is inserted before the saturation; it is necessary to eliminate the zero by expanding the transfer function $(\alpha Ts+1)/(Ts+1)$ into partial fractions, as shown in Fig. 12.

If the co-ordinates in $\mathfrak{U}$ are again defined by Fig. 2(B), the transformation from $\mathbb{C}$ to $\mathfrak{U}$ is

$$c=P_u u=\begin{Vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & T & -T^2 \end{Vmatrix}u \qquad (31)$$

The saturation-region trajectories in the $(r-u_1, \dot{r}-u_2)$ phase plane are parabolas as in examples 1 and 3.

The root loci are the same as in the previous example, with αT replacing K_2. The co-ordinates in $\mathfrak{V}$ being again defined by Fig. 9(B), the transformation from $\mathbb{C}$ to $\mathfrak{V}$ is given by

$v=P_v^{-1}c=$

$$\begin{Vmatrix} 1 & T_r-\alpha T & -\alpha T(T_r-\alpha T)K_1 \\ 0 & T_c & T_c(T_r-\alpha T)K_1 \\ 1 & 2\zeta T_c-\alpha T & [T_c^2+\alpha T(\alpha T-2\zeta T_c)]K_1 \end{Vmatrix}c$$

It is readily verified from equation 31 and Fig. 12 that the eigenvector equation in $\mathfrak{U}$ is again the same as in $\mathbb{C}$

$s_i(r-u_1)=\dot{r}-u_2$ (32)

When entered from saturation, the boundaries of the linear region are

$(r-u_1)+(\alpha-1)T(\dot{r}-u_2)$
$\quad =\pm[1/K_1+T^2(1-\alpha)]+(\alpha-1)\dot{r}$ (33)

More interesting are the boundaries of the linear region for trajectories leaving it

$[\alpha-1+\alpha(T/T_r+\alpha T T_rK_1)](r-u_1)-$
$\quad (\alpha-1)(\dot{r}-u_2)=\mp(1/K_1+\alpha T_r^2)T/T_r+$
$\quad\quad\quad\quad\quad\quad (1-\alpha)\dot{r}$ (34)

It is now easy to run down the various cases of interest as far as the variation of both K_1 and α is concerned:

$\alpha>9$. If $0<K_1<K_{11}$, the transient response in the linear region is too slow. Increasing the gain to $K_{11}\leq K_1\leq K_{12}$ makes the dominant eigenvalues real; by equations 32

through 34 they always remain inside the linear region and therefore the transient never overshoots regardless of the magnitude of the input. If $K_1 > K_{12}$, not only does the linear region contract but the spirals corresponding to the complex eigenvalues may leave it. Hence $K_1 = K_{12}$ is a practical limit on the gain. Note also that a large α tends to slow down the response in the linear region.

$1 < \alpha < 9$. No matter how small $K_1 > 0$ is chosen, the spiral trajectories corresponding to the dominant complex eigenvalues will not lie entirely within the linear region. This imposes a limitation on large signal performance since beyond a certain input magnitude the trajectories will overshoot. On the other hand, reduction of α permits somewhat faster response in the linear region.

$\alpha < 1$. This corresponds to a lag network; the system is totally unstable.

EXAMPLE 6. ROLL-CONTROL OF MISSILE

As a concluding example, consider a missile roll-control problem where the influence of saturation is not immediately clear from physical reasoning. The aero-

dynamic relationship between the control surface deflection c_3 and the actual roll angle c_1 is given approximately by the transfer function $1/s(Ts+1)$; in turn, c_3 is governed by a high-gain hydraulic servomechanism whose input is proportional to the roll-angle error e. To conserve hydraulic fluid, the servomechanism must be velocity-limited; this effect is the nonlinearity to be analyzed. The signal-flow graph is shown in Fig. 13(A).

The transformation from $\mathbb{C}$ to $\mathfrak{U}$ is similar to that used in examples 3 and 4; it is

$$ c = P_u u = \begin{Vmatrix} 1 & -T & T^2 \\ 0 & 1 & -T \\ 0 & 1 & 0 \end{Vmatrix} u $$

provided the co-ordinates in $\mathfrak{U}$ are again selected in accordance with Fig. 2(B).

Since this is a regulating system, it is proper to use the (u_1, u_2) phase plane where initial conditions correspond to disturbances which the system is designed to counteract. The boundaries of the linear region for trajectories entering from saturation are

$$ K_3 u_1 + (1 - K_3 T) u_2 = \mp (1/K_1 + K_3 T^2) $$

Assume now that $K_1 \gg 1/T$, as is generally the case in practice. The linear zone is then very narrow and, for large inputs, the system behaves as a relay servo in the (u_1, u_2) phase plane with an open-loop transfer function $1/s^2$ and a reversal line of slope $T - 1/K_3$. If the slope is positive, i.e.

$K_3 > 1/T$

the relay servo may be made completely unstable by a large, sudden disturbance. However, unlike in examples 4 and 5, the foregoing inequality does not imply instability in the linear region. Since $1/K_1 \cong 0$, the eigenvalue equation is approximately

$$ s^2 + s/T + K_3/T = 0 \tag{35} $$

Setting $K_3 = 1/T$ leads to a damping ratio $\zeta = 0.5$ for the eigenvalues of equation 35. Even for any $K_3 > 1/T$, the transients in the linear region will be stable, provided that K_1 is sufficiently large. Thus small trajectories will converge to the origin while large trajectories diverge to infinity, revealing the presence of an unstable limit cycle. The limit cycle will be composed of half-parabolas and may be calculated exactly in the (u_1, u_2) phase plane by the graphical construction shown in Fig. 13(B). The reversal curve bends toward the origin as indicated, since $|u_3| < 1$ if the reversal takes place near the origin.

The unstable limit cycle defines a stability boundary in $\mathfrak{U}$ as follows: Move the origin of the (u_1, u_2) plane along the u_3 axis, maintaining the orthogonality of the axes. The limit cycle will generate the surface of a cylinder. All trajectories starting within the cylinder will converge to the origin, and all others will diverge to infinity. For very large values of u_3, the cylinder will become distorted since the approximation $u_3 \cong \pm 1$ used in deriving the reversal point will no longer be good.

(A)

(B)

Fig. 13. Example 6

A—Signal-flow graph in $\mathbb{C}$
B—Determination of limit cycle
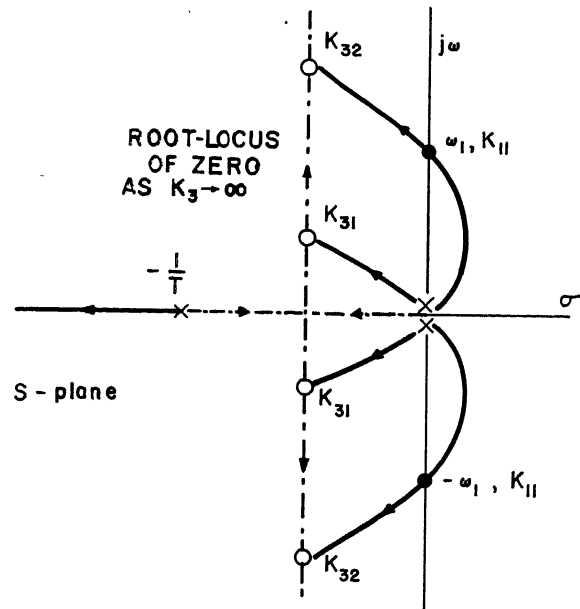
Fig. 14 (above). Example 6; root locus as a function of K_1, with K_3 as parameter

Comparison With Describing-Function Method

It is interesting to consider the question of establishing the existence and of calculating the approximate period and amplitude of the unstable limit cycle by the describing-function method. Johnson[34] has calculated the amplitude and phase of the fundamental harmonic component of c_3 as a function of the amplitude and frequency of a sinusoidal forcing signal e, thereby obtaining a describing function for the velocity-limited servomechanism in Fig. 13(A). He, too, concludes that an unstable limit cycle exists for sufficiently large K_3. An equivalent but conceptually simpler way of proceeding is this.[35] Write the eigenvalue equation for the linear region of the system as

$$K_1 \frac{s(Ts+1)+K_3}{s^2(Ts+1)}+1=0 \qquad (36)$$

Equation 36 represents a root locus as a function of K_1. The zeros of the root locus in turn depend on K_3, as shown by the dash-dotted line in Fig. 14. For nonlinear operation, K_1 is to be interpreted as some equivalent gain. Thus K_1 will be very large for linear operation (x small) and approach zero for highly nonlinear operation (x large). Now if K_3 is fairly small (K_{31}), the root locus will not cross over into the right-half plane, and the system will be stable for all values of K_1. If K_3 is sufficiently large (K_{32}), however, the root locus will cross and instability will result whenever K_1 is small. Hence, there will be an unstable limit cycle, with its approximate amplitude and frequency being given by the intersection of the root locus with the $j\omega$ axis, as in Fig. 14.

Notice that the foregoing arguments depend crucially on what is meant by equivalent gain. If the a priori assumption is made that a stable or unstable limit cycle exists, then K_1 may be defined as the ratio of the first-harmonic components in $\dot{c}_3 = f(x)$ and x, justified by noting that the higher harmonics are attenuated by the transfer function between $\dot{c}_3$ and x. If the limit cycle actually exists and is approximately sinusoidal, the root locus must necessarily intersect the $j\omega$ axis. But intersection of the axis is not sufficient to infer the existence of a limit cycle. Thus the success of the describing-function method may depend on whether a limit cycle actually exists—the very fact that the method is expected to prove. In fact, Nichols[36] has pointed out that the describing-function method applied to a backlash problem indicates the presence of an unstable limit cycle whereas exact analysis shows that none exists. Despite

much recent effort,[37] this fundamental weakness of the describing-function method has not been overcome.

By contrast, the phase-space analysis outlined in the foregoing is not dependent on any a priori assumption. The only approximation, concerning u_3, is made in the time domain and is very easily checked in any particular case.

Conclusions

All of the examples of this section were governed by the same saturation-region differential equations, and yet the effect of saturation was quite different in each case. These differences were caused by the manner in which saturation was included in the systems; they were revealed by a detailed consideration of how the trajectories entered and left the linear region.

Summary

This paper presents, apparently for the first time in the engineering literature, a simple approach to high-order nonlinear systems based on linear transformations in the phase space and closely linked to the very useful transfer-function concept in the theory of linear transients. In the present form of the theory, detailed quantitative calculations may be quite laborious. But only a simple analysis of the type discussed is needed to discover characteristic features of the nonlinear system; for this task, the phase-space method is admirably suited.

The author is convinced that the combination of the phase-space representation of nonlinear differential equations, linear transformations linked to the familiar partial-fraction expansion of transfer functions, and the root-locus method of stability study together provide a powerful and modern mathematical apparatus for the study of nonlinear control problems. Much further progress may be expected from the exploitation of these ideas.

Appendix I. Trajectories of Second-Order Linear Systems

For quantitative work in phase-plane analysis, it is very useful to have a set of standard curves of the response of second-order linear systems. A series of such scale plots are reproduced here for ready reference; see Fig. 15. All curves were obtained by means of a real-time analogue computer; accuracy is about ±2 per cent.

The curves represent solutions of the equation often used in servo work

$$\frac{d^2x}{dt^2}+2\zeta\frac{dx}{dt}+x=0 \qquad (37)$$

where ζ is a real parameter, the viscous damping coefficient. The qualitative features of Fig. 15 are best discussed in terms of the eigenvalues s_i of equation 37. In view of the diagonalization theorem, this will also take care of the general case where A is a 2×2 matrix.

Case 1. Stable Node: $1 \leq \zeta < \infty$; or s_1, s_2 are real, negative. The solutions in normal co-ordinates are $u_i(t)=u_i(0)\epsilon^{s_i t}$ ($i=1$, 2). From these, the equation of the trajectories is readily obtained by eliminating the time

$$u_2 = \text{constant } u_1^{s_2/s_1} \quad s_2 < s_1 < 0 \qquad (38)$$

Equation 38 represents a family of parabolas symmetrical around the u_2 axis. Transformation to the $(x, \dot{x})$ phase plane distorts the trajectories, but does not affect the nature of the critical point. Thus in Figs. 15(A) and (C) the trajectories start out as straight lines parallel to the eigenvector corresponding to s_2 (the u_2 axis), and they approach asymptotically the eigenvector corresponding to s_1 (the u_1 axis). This means that the initial rise time of a 2-lag system in response to a step depends primarily on the smaller s_i whereas the final settling time is governed by the larger s_i, as is well known. Bearing these simple physical notions in mind, it is very natural to approximate the trajectories in Fig. 15(A) by two straight lines, as shown in Fig. 15(B), one running parallel to the u_2 axis (the "fast" eigenvector) and the other being the u_1 axis itself (the "slow" eigenvector). Because of the much smaller separation between the eigenvalues, the approximation is less accurate if applied to Fig. 15(C).

Case 2. Stable Focus: $0 < \zeta < 1$ or s_1, s_2 are complex conjugate, $\sigma < 0$. Here it is best to regard one of the solutions in normal co-ordinates

$$u_1(t)=u_1(0)\epsilon^{s_1 t}=u_1(0)\epsilon^{(a+jb)t} \qquad (39)$$

as a first-order differential equation defined over a 1-dimensional complex space, i.e., the ordinary complex plane. The other solution is merely the complex conjugate of equation 39. The complex quantity $u_1(t)$ is to be regarded as a rotating vector with bt being the angle of rotation in radians and ϵ^{at} the attenuation of the magnitude of the vector. The curve described by the tip of the vector, a logarithmic spiral, is the trajectory in the $(Re[u_1], Im[u_1])$ plane. When translated into the $(x, \dot{x})$ phase plane, the logarithmic spirals appear as in Figs. 15(D), (E), and (F).

It is unfortunately not possible to make use here of the asymptotic behavior of the trajectories as in the case of the node; however, explicit computation of the trajectory is not too difficult in view of the convenient form of equation 39.

Case 3. Center: $\zeta = 0$, or s_1, s_2 are pure imaginary. In equation 39, a is set to zero and therefore the rotating vectors describe a circle. The trajectories consist of concentric circles. The system is said to be conservative.

The qualitative difference between parabolic, spiral, and circular trajectories is indeed striking; this is the motivation for

$\ddot{x}+2\zeta\dot{x}+x=0$
$\zeta = 2.0$

(A)

$\ddot{x}+2\zeta\dot{x}+x=0$
$\zeta = 0.7$

(D)

"FAST" EIGENVECTOR

"SLOW" EIGEN-VECTOR

"SLOW" EIGEN-VECTOR
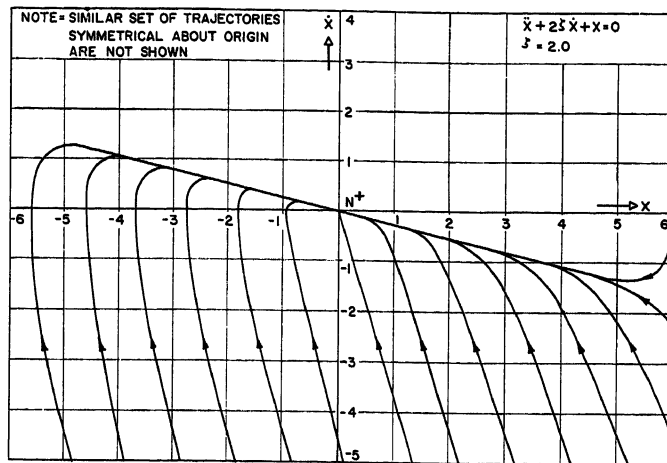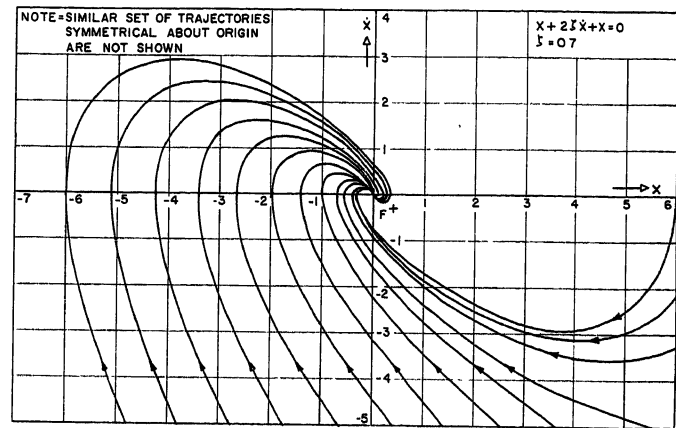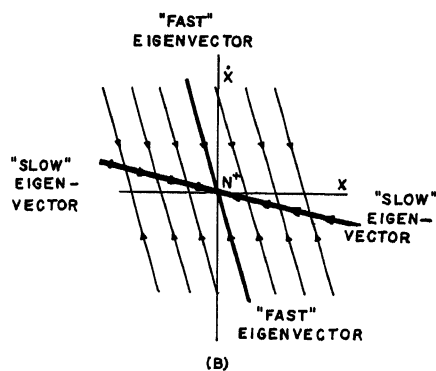
"FAST" EIGENVECTOR

(B)

Fig. 15. Trajectories of the second-order linear differential equation

$$\frac{d^2x}{dt^2}+2\zeta\frac{dx}{dt}+x=0$$

A—$\zeta=2.0$
B—Approximate representation with the help of eigenvectors
C—$\zeta=1.2$
D—$\zeta=0.7$
E—$\zeta=0.4$
F—$\zeta=0.2$

$\ddot{x}+2\zeta\dot{x}+x=0$
$\zeta = 0.4$

(E)

$\ddot{x}+2\zeta\dot{x}+x=0$
$\zeta = 1.2$

(C)

$\ddot{x}+2\zeta\dot{x}+x=0$
$\zeta = 0.2$
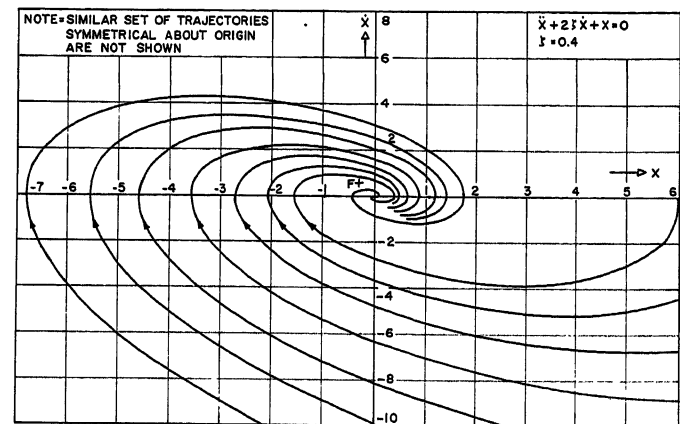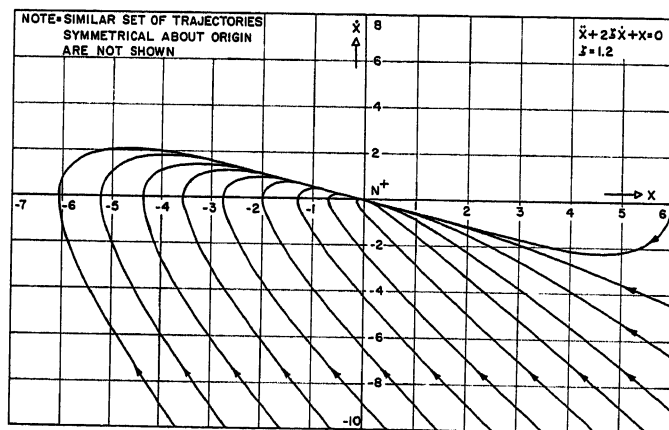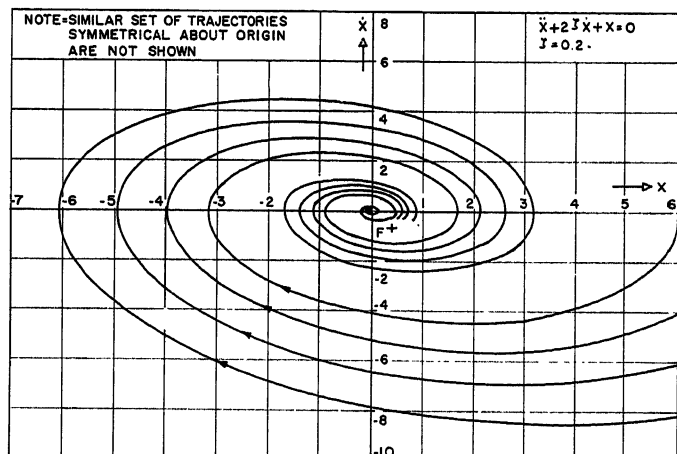
(F)

the insistence in phase-plane theory on sharply differentiating node, focus, and center type of critical points and trajectories.

The case of the saddle point, i.e., s_1, s_2 real, $s_1<0<s_2$, is not treated; it has not been important up to the present in servomechanism theory. Since the eigenvalues are real, the trajectories are easily visualized using the eigenvector concept, as in the case of a node. The degenerate cases when one or both eigenvalues vanish have been treated at some length in connection with optimum relay servos.[3-6,9]

All the trajectories in Fig. 15 represent stable systems, i.e., the trajectories tend to be the critical point as $t \to +\infty$; hence the superscript + over the abbreviation of the critical point.[18] Unstable trajectories may

be obtained by reflection about the $\dot{x}$ axis, which corresponds to changing the sign of ζ.

Appendix II. Optimum Relay Servomechanism Theory

Second-Order Systems

The problem of how a relay-controlled or saturating servomechanism is to be designed for optimum speed of response may be completely solved by use of the phase-space approach. The correct solution may be discovered by the following intuitive argument:

To have optimally fast response, it is necessary to exploit fully the power (or torque, etc.) capabilities of the system; hence, the signal into the output element must be maintained at either forward or reverse saturation, corresponding to the two nonzero positions of a polarized relay. This suggests that the signal (forcing function) applied to the output element should be idealized as having only three distinct values: $+1$, -1, and 0. There will be two different families of trajectories in the phase plane corresponding to the two signs. The problem is to bring the system to rest at zero error as quickly as possible. In the $(e, \dot{e})$ phase plane this is achieved by leading the trajectory into the origin.

Clearly, the transient can reach the origin only by following a trajectory which goes through the origin. There exists one and only one such trajectory in each family since the origin is not a critical point. These trajectories are called reversal curves. They subdivide the phase plane into two regions. In each region the signal polarity is so specified that the resulting trajectories will always intersect the reversal curve, at which instant the signal polarity is reversed. The transient then proceeds along the reversal curve. At the instant of reaching the origin, the signal to the output element is set to zero and the system remains at rest.

This scheme was first proposed and demonstrated with simple examples by McDonald[3] and Hopkin[5] and many other contributions have been recorded since then.[6,15,38-43] In a doctorate thesis, Bushaw[42] has rigorously proved that the intuition which led to the formulation of the theory is indeed correct.

In view of the significance of the optimum relay servomechanism concept in the analysis of saturating systems, the various important special cases of second-order transfer functions are now briefly reviewed. A detailed discussion is given by Tsien.[44]

Case 1. $G(s)=1/s^2$, e.g., ideal torque source driving pure inertia.[3,39] The trajectories are families of parabolas, with vertexes on and symmetrical with the e axis. Representative transient trajectories are sketched in Fig. 16(A). Regions in the phase plane are labeled with $+$ and $-$ signs according to the signal polarity required. The equation of the reversal curve is

$$2e = -\dot{e}|\dot{e}| \tag{40}$$

Case 2. $G(s)=1/s(s+1)$, e.g., armature-controlled d-c motor.[5,39] For small velocities, the trajectories are similar to case 1. Ultimately they become asymptotic to the dashed lines $\dot{e}=\pm1$; see Fig. 16(B). Note that the trajectories may approach the dotted lines also from the outside. The equation of the reversal curve is

$$\frac{e}{T} = \dot{e}\left(1 - \frac{ln\,(1+|\dot{e}|)}{|\dot{e}|}\right) \tag{41}$$

Case 3. $G(s)=1/(s^2+2\zeta s+1)$, $\zeta\geq1$, e.g., 2-capacity temperature regulation problem.[42] The trajectories tend to the stable nodes $(N_-{}^+)$ and $(N_+{}^+)$, depending on whether they originate above or below the reversal curve; see Fig. 16(C). Only the arc $A_+ O A_-$ on the reversal curve is of practical interest, since all large-signal trajectories having converged to the slow eigenvector of $(N_+{}^+)$ will intersect the reversal line at A_+. Similary for A_-.

Case 4. $G(s)=1/(s^2+2\zeta s+1)$, $\zeta\leq-1$, unstable open loop. The trajectories belong to the unstable nodes $(N_+{}^-)$ and $(N_-{}^-)$ respectively. Connecting $(N_+{}^-)$ with trajectory Γ_- to $(N_-{}^-)$ and $(N_-{}^-)$ with Γ_+ to $(N_+{}^-)$, Fig. 16(D), it is easily seen that all trajectories originating outside the closed curve $\Gamma_+\Gamma_-$ will diverge to infinity. If the reversal curve is again made up of trajectories going through the origin and terminating at the nodes, stable operation is assured everywhere inside the curve $\Gamma_+\Gamma_-$. For any other choice of the reversal curve, say, a straight line through the

origin realizable by linear velocity feedback, the stable region will be necessarily smaller. The boundary $\Gamma_+\Gamma_-$ is analogous to an unstable limit cycle.

In the foregoing cases, optimum control is achieved by a total of two polarity reversals, including in the count the fact that the signal must be set to zero on reaching the origin. The case of a second-order transfer function with complex roots is in a sense exceptional since only a part of the optimum reversal curve is a trajectory; more and more reversals are necessary as the magnitude of the initial error is increased.

Case 5. $G(s)=1/(s^2+2\zeta s+1)$, $0<\zeta<1$, e.g., aircraft regulation problem.[42] The trajectories are spirals, belonging to the stable foci $(F_+{}^+)$ and $(F_-{}^+)$. The two trajectories leading into the origin intersect one another an infinite number of times; moreover, they do not subdivide the phase plane into two distinct parts. The problem requires serious mathematical investigation.

According to Bushaw's solution,[42] the reversal curve consists of arcs of trajectories constructed as follows: Take a trajectory belonging to $(F_+{}^+)$ and follow it backwards in time from the origin until it intersects the c axis at c_p where $p=1$. Now magnify the arc so obtained by the factor $\exp$ $p(\pi\zeta/\sqrt{1-\zeta^2})$ and displace it; c_p units to the right, continuing the process with $p=1, 2, 3, \ldots$. In other words, that trajectory of $(F_-{}^+)$ which goes through the origin is cut at each intersection with the c axis, and the arcs so obtained are placed side by side. The reversal curve for $c<0$

Fig. 16. Reversal curves for optimally compensated second-order systems classified according to output element transfer function

A—$1/s^2$
B—$1/s(Ts+1)$
C—$1/(s^2+2\zeta s+1)$, $\zeta>1$
D—$1/(s^2+2\zeta s+1)$, $\zeta<-1$
E—$1/(s^2+2\zeta s+1)$, $0<\zeta<1$

is determined similarly; see Fig. 16(E). If the initial point lies on the c axis between $c_{p-1}<|c|<c_p$, the number of reversals will be $p+1\geq2$.

It is difficult to approach optimum performance in this case by linear compensation. A thorough treatment of the many phenomena encountered in relay servos with velocity feedback and this transfer function is given by Flügge-Lotz.[10]

REMARK ON INPUTS

Because of considerations of steady-state error, if the locations of the optimum reversal curves are to be independent of the input, only the following types of inputs are admissible to servomechanisms (follow-up systems) incorporating the transfer functions discussed in the foregoing:

Case 1. Steps and/or ramps.
Case 2. Steps only.
Cases 3 through 5. Neither steps nor ramps.

The axes in Fig. 16 have been labeled accordingly.

Since arbitrary inputs may be approximated by means of steps and ramps, the response of systems of the type 1 and 2 may be regarded as generally acceptable though not always optimum, provided that in the steady state the inputs behave as listed in the foregoing. In cases 3 through 5, only inputs which are identically zero in the steady state are allowed; hence these cases are of interest mainly in connection with regulating systems.

On the other hand, system 1 is capable of following both step and ramp inputs with zero steady-state errors and nonovershooting transients. This contrasts with the well-known fact that a linear servo which

has zero steady-state error to a ramp input necessarily overshoots in response to a step.

Remark on Zeros

A further unexpected difficulty appears when the transfer function is allowed to have a zero. The reversal curve, composed of the trajectories leading into the origin of the normal co-ordinate phase plane, depends only on the poles and not on the zero. On the other hand, the transformation which takes the reversal curve back to the $(e, \dot{e})$ phase plane is a function of the zero. By proper selection of the zero, the reversal curve may be rotated by an arbitrary angle in $(e, \dot{e})$; in particular, it may be made to intersect the line $e=0$ so that the transient will overshoot once. Under such circumstances, the response cannot be regarded as optimum and it may be mandatory to introduce a linear zone. The foregoing statements are easily checked by considering an optimum relay servo with the transfer function $(Ts+1)/s^2$, which will have an overshooting step response if $T>0$. These remarks should emphasize the fact that, while an optimum response may always be defined, the existence of meaningful optimum system depends on both the nature of the transfer function and of the inputs.

Systems of Arbitrary Order

If the concept of second-order optimum relay servomechanisms is thoroughly understood, there is no basic difficulty in extending it to arbitrarily high-order systems. Because of the unpleasant complications which have arisen in connection with case 5, only open-loop transfer functions with real poles have been considered so far.[12,14,43]

The argument proceeds backwards: In n-dimensional phase space, the $\pm$ trajectories through the origin constitute the reversal curve. Each point on this reversal curve may be reached (is intersected) by a trajectory of the opposite polarity; the latter trajectories generate a 2-dimensional reversal surface. This process may be continued until the trajectories intersecting the last $(n-1)$-dimensional reversal surface fill the entire n-dimensional phase space. Hence n reversals are necessary in the n-dimensional case. Take a third-order system such as in examples 3 and 4. The first reversal occurs when the trajectory intersects the 2-dimensional surface; after that, the trajectory remains in that surface until intersecting the trajectory through the origin, when the polarity is again reversed. By definition, the third reversal occurs when the signal becomes zero on reaching the origin.

The equations describing the reversal surface may be readily obtained by means of the normal co-ordinate transformation. The problem has been worked out for the transfer function $1/s^2(Ts+1)$ by Bogner.[12] A more general treatment is due to Rose.[43]

The incorporation of a reversal surface in a practical system is cumbersome because of the necessity of simulating the 1, 2, ... $(n-1)$-dimensional reversal surfaces. Moreover, it would be necessary to measure accurately the first $(n-1)$ derivatives of the error. Herein lies the importance of the approximation scheme discussed in the foregoing. Still, in a difficult control problem, crude simulation

of a reversal surface by a diode network may well be the only adequate method of compensation.

Notice also that the optimum compensation assumes precise knowledge of an unchangeable transfer function. The optimum system is a calibrated system. Much further work is necessary before practical equipment can be built which will reliably approach optimum performance even when subjected to noise, slight changes in parameters, etc.

References

1. Fundamental Theory of Servomechanisms (book), L. A. MacColl. D. Van Nostrand Company, Inc., New York, N. Y., 1945, Appendix.

2. Analysis of Relay Servomechanisms, H. K. Weiss. *Journal of the Aeronautical Sciences*, Easton, Pa., vol. 13, 1946, pp. 364–76.

3. Nonlinear Techniques for Improving Servo Performance, D. C. McDonald. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 6, 1950, pp. 400–21.

4. Multiple-Mode Operation of Servomechanisms, D. C. McDonald. *Review of Scientific Instruments*, New York, N. Y., vol. 23, 1952, pp. 22–30.

5. A Phase-Plane Approach to the Design of Saturating Servomechanisms, A. M. Hopkin. *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 631–39.

6. Automatic and Manual Control (book). Academic Press, New York, N. Y., 1952, "The Stablization of On-Off Controlled Servomechanisms," A. M. Uttley, P. H. Hammond, pp. 285–308.

7. Analysis of a Friction Damper for Clutch-Type Servomechanisms, H. K. Weiss. *Journal of the Aeronautical Sciences*, Easton, Pa., vol. 18, 1951, pp. 676–82.

8. Graphical Solution of Some Automatic-Control Problems Involving Saturation Effects with Application to Yaw Dampers for Aircraft, W. H. Phillips. *NACA Technical Note No. 3034*, National Advisory Committee for Aeronautics, Washington, D. C., 1953.

9. The Application of Nonlinear Techniques to Servomechanisms, K. C. Mathews, R. C. Boe. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 8, 1952, pp. 10–21.

10. Discontinuous Automatic Control (book), I. Flügge-Lotz. Princeton University Press, Princeton, N. J., 1953.

11. Design Considerations of a Saturating Servomechanism, P. E. Kendall, J. F. Marquardt. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 9, 1953, pp. 178–87.

12. An Investigation of Switching Criteria for Higher Order Contactor Servomechanisms, I. Bogner, L. F. Kazda. *AIEE Transactions*, vol. 73, pt. II, July 1954, pp. 118–27.

13. Phase-Plane Analysis of Automatic Control Systems with Nonlinear Gain Elements, R. E. Kalman. *Ibid.*, Jan. 1954, pp. 383–90.

14. Nonlinear Control of a Saturating Third-Order Servomechanism, J. L. Preston. *M. S. Thesis*, Massachusetts Institute of Technology, Cambridge, Mass., May 1954.

15. The Effects of the Addition of Some Nonlinear Elements on the Transient Performance of a Simple R. P. C. System Possessing Torque Limitation, J. C. West, J. L. Douce, R. Naylor. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 101, pt. II, 1954, pp. 156–65.

16. The Step-Function Response of an R. P. C. Servo Mechanism Possessing Torque Limitation, J. C. West, I. R. Dalton. *Ibid.*, pp. 166–73.

17. Theory of Oscillations (book), A. A. Andronow, C. E. Chaikin. Princeton University Press, Princeton, N. J., 1949.

18. Introduction to Nonlinear Mechanics (book), N. Minorsky. J. W. Edwards, Ann Arbor, Mich., 1947.

19. Geometrical Methods in the Analysis of Ordinary Differential Equations: Intro-

duction to Nonlinear Mechanics, J. Kestin, S. K. Zaremba. *Applied Scientific Research*, The Hague, Holland, vol. 3B, 1953, pp. 149–89.

20. Synthesis of Feedback Control Systems by Phase-Angle Loci, Y. Chu. *AIEE Transactions*, vol. 71, pt. II, 1952, pp. 330–39.

21. The Study of Transients in Linear Feedback Systems by Conformal Mapping and the Root-Locus Method, V. C. M. Yeh. *Transactions*, American Society of Mechanical Engineers, New York, N. Y., vol. 76, 1954, pp. 349–61.

22. Lectures on Differential Equations (book), S. Lefschetz. Princeton University Press, Princeton, N. J., 1946.

23. Stability Theory of Differential Equations (book), R. Bellman. McGraw-Hill Book Company, Inc., New York, N. Y., 1953.

24. Finite Dimensional Vector Spaces (book), P. R. Halmos. Princeton University Press, Princeton, N. J., 1942.

25. A Survey of Modern Algebra (book), G. Birkhoff, S. MacLane. The Macmillan Company, New York, N. Y., 1946.

26. Feedback Theory. Some Properties of Signal Flow Graphs, S. J. Mason. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 41, 1953, pp. 1144–56.

27. Automatic Feedback Control System Synthesis (book), J. A. Truxal. McGraw-Hill Book Company, Inc., New York, N. Y., 1955.

28. Electronic Analog Computers (book), G. A. Korn, T. M. Korn. *Ibid.*, 1952.

29. Introductory Circuit Theory (book), E. A. Guillemin. John Wiley and Sons, Inc., New York, N. Y., 1953.

30. Transient Vibration of Linear Multi-Degree-of-Freedom Systems by the Phase-Plane Method, R. S. Ayre. *Journal*, Franklin Institute, Philadelphia, Pa., vol. 253, 1952, pp. 153–66.

31. Metrization of Phase Space and Nonlinear Servo Systems, Chi Lung Kang, G. H. Fett. *Journal of Applied Physics*, New York, N. Y., vol. 25, 1955, pp. 38–41.

32. Calculation of the Inverse of a Matrix, R. V. Andree. *American Mathematical Monthly*, Buffalo, N. Y., vol. 58, 1951, p. 87.

33. The Synthesis of "Optimum" Transient Response: Criteria and Standard Forms, D. Graham, R. C. Lathrop. *AIEE Transactions*, vol. 72, pt. II, Nov. 1953, pp. 273–88.

34. Proceedings of the Symposium on Nonlinear Circuit Analysis (book). Polytechnic Institute of Brooklyn, Brooklyn, N. Y., 1953, "Sinusoidal Techniques Applied to Nonlinear Feedback Systems," E. C. Johnson, pp. 258–73.

35. *Discussion* of H. Chestnut's paper, R. E. Kalman. *Transactions*, American Society of Mechanical Engineers, New York, N. Y., vol. 76, 1954, p. 1362.

36. Backlash in a Velocity Lag Servomechanism, N. B. Nichols. *AIEE Transactions*, vol. 72, pt. II, Jan. 1953, pp. 462–67.

37. Recent Advances in Nonlinear Servo Theory, J. M. Loeb. *Transactions*, American Society of Mechanical Engineers, New York, N. Y., vol. 76, 1954, pp. 1281–89.

38. On the Comparison of Linear and Nonlinear Servomechanism Response, T. M. Stout. *Transactions*, IRE Professional Group for Circuit Theory, Institute of Radio Engineers, New York, N. Y., vol. CT-1, Mar. 1954, pp. 49–55.

39. Effects of Friction in an Optimum Relay Servomechanism, T. M. Stout. *AIEE Transactions*, vol. 72, pt. II, Nov. 1953, pp. 329–36.

40. Switching Errors in an Optimum Relay Servomechanism, T. M. Stout. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 9, 1953, pp. 188–98.

41. Predictor Servomechanisms, L. M. Silva. *Transactions*, IRE Professional Group for Circuit Theory, Institute of Radio Engineers, New York, N. Y., vol. CT-1, Mar. 1954, pp. 56–70.

42. Differential Equations with a Discontinuous Forcing Term, D. W. Bushaw. *Ph.D. Thesis*, Princeton University, Princeton, N. J., 1952.

43. Theoretical Aspects of Limit Control, N. J. Rose. *Report No. 459*, Stevens Institute of Technology, Hoboken, N. J., Nov. 1953.

44. Engineering Cybernetics (book), H. S. Tsien. McGraw-Hill Book Company, Inc., New York, N. Y., 1954.

Discussion

J. M. Loeb (Schlumberger Instrument Company, Ridgefield, Conn.): The author should be commended for having given a rigorous approach to an approximation method used by some other authors. Since higher order nonlinear systems are dealt with, a constant need for a kind of "phase plane approach" has been experienced.

The existence of computers has changed the aim of theoretical calculations. For all particular problems, i.e., derivation of the solution of a set of nonlinear differential equations being given the necessary limit conditions, no special mathematical concept is necessary. A good analogue or digital computer does the job. As a matter of fact, the phase plane approach, introduced by Cauchy to drive his "unicity theorem," is rarely used for practical computation. However, the control engineer has equally to perform synthesis of equipments and a general visualization becomes necessary.

Mr. Kalman has showed very well how matrix algebra leads to the discovery of independant variables particularly suitable for a phase plane approximation, i.e., that are related by differential equations containing largest time constants. The conception of "projection" of the phase space upon this particular plane is the key to a good understanding of a whole class of interesting systems.

It remains, however, very difficult to cross the gap that separates linear problems from nonlinear ones. All the problems treated in the paper belong to the so-called "piecewise linear" field.

Many systems are devised to supply in due time the best "switching" that changes the configuration of the system. In more complicated diagrams, is it not to be foreseen that the decomposition, or the author's "near-diagonalization," would lead to a different couple of principal variables? If it is so, one would have to operate the inverse transformation each time the representing point crosses the boundary between different configurations of the system.

K. Klotter (Stanford University, Stanford, Calif.): I share the author's opinion (as expressed in the summary) that the "combination of the phase-space representation of nonlinear differential equations, linear transformations linked to the familiar partial-fraction expansion of transfer functions, and the root locus method of stability study together provide a powerful . . . . apparatus for the study of nonlinear control problems," and think that the author did a commendable job in presenting that technique of investigation.

However, in order to keep the limitations of the procedure in the reader's mind, I would like to underline, and also to implement, a few statements in the paper which are apt to become lost in the technical details:

1. The nonlinear systems investigated are replaced by piecewise linear ones. Although it is true that such an approximation may be achieved "with arbitrary accuracy," it is equally true that increased accuracy entails highly increased labor.

2. The driving terms are replaced by constants; they, in turn, by a transformation of co-ordinates, can be eliminated completely. In this way homogeneous linear equations result.

3. The system of nth order finally is replaced by one of second order. Hence, intrinsically, no "extension of the phase plane method to high-order nonlinear differential equations" is involved. The method is as good as the replacement is. Furthermore, all examples given are restricted to third order.

The author, in presenting his case, makes extensive use of the parlance within some circles of communication engineers, including the use of signal flow graphs in lieu of differential equations. In using, and even coining, new expressions for familiar algebraic and analytical concepts, the author at one point, however, transcends the permissible limits: In the paragraph following equation 1 the quantities c_i and c_t are referred to as "dependent variables" and the quantities r_t as "independent variables." Such a usage is at variance with well-established mathematical terminology. The c_t are the dependent variables; however, the r_t are the "driving terms" (or "forcing terms" or "perturbation terms," etc.) but certainly not the independent variables. Independent variable is the time t exclusively and throughout.

Y. H. Ku (Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa.): This paper is a serious attempt to generalize phase-plane methods of dealing with nonlinear systems and appears to pinpoint a promising area of future research. To represent the solutions of nonlinear differential equations of the nth order, it is necessary to use an n-dimensional phase space, in which an n-dimensional space trajectory representing the solution is located.[1,2] The author is to be congratulated for giving a simplified method of approximating the actual trajectory by breaking up the phase space into a number of regions, such that within each region the portions of the approximated trajectory represent the solutions of linear differential equations. However, it may be pointed out that in the general phase-space method, it is not absolutely necessary that such linear approximations are made. It is a straightforward procedure to find the actual trajectory if it needs be found.[3,4] Thus the phase space of n-dimensions offers a seat to the actual trajectory as well as to the approximated portions of the actual trajectory. In fact, while the junctions of the approximated portions have multivalued slopes, the actual trajectory has continuity of slopes at all points. It would be of interest to compare the actual trajectory (with nonlinearities as originally specified or experimentally determined) with the portions of the approximated trajectory. The simplified method would be especially useful in synthesis and design of nonlinear servomechanisms.

References

1. A METHOD FOR SOLVING THIRD AND HIGHER ORDER NONLINEAR DIFFERENTIAL EQUATIONS, Y. H. Ku. *Journal*, Franklin Institute, Philadelphia, Pa., vol. 256, no. 3, Sept. 1953, pp. 229–44.

2. ANALYSIS OF NONLINEAR SYSTEMS WITH MORE THAN ONE DEGREE OF FREEDOM BY MEANS OF SPACE TRAJECTORIES, Y. H. Ku. *Ibid.*, vol. 259, no. 2, Feb. 1955, pp. 115–31.

3. ANALYSIS OF NONLINEAR COUPLED CIRCUITS. Y. H. Ku. *AIEE Transactions*, vol. 73, pt. I, 1954 (Jan. 1955 section), pp. 626–31.

4. PART II. *Ibid.*, vol. 74, pt. I, Sept. 1955, pp. 439–43.

Herbert K. Weiss (Northrop Aircraft, Inc., Hawthorne, Calif.): This paper represents a definite contribution to the understanding of nonlinear systems. The brief section on co-ordinate transformations and the several examples are particularly helpful.

Since most interesting nonlinearities can be adequately represented by the saturation function shown in Fig. 3 of the paper, the methods developed appear to be of considerable generality. It is not clear, however, that examination of the latter phases of the transient motion only, as accomplished by use of the two dominant eigenvalues, will yield a sufficiently complete picture for all servo analyses. This is a question of the value function by which one defines the performance objectives to be sought in the servo design. One might imagine, for example, a servo subjected to continued and rapid disturbances, or one attempting to follow a signal in the presence of rapidly fluctuating noise. In these cases the long period response to a single step or ramp input might be of minor importance compared with the initial response gradient to signal change.

It is hoped that Mr. Kalman will find time to extend the methods of his present paper to the examination of all stages of servo response.

R. E. Kalman: I wish to thank the discussers for their serious interest in the paper. Indeed, most of the comments deal not with the specific results of the paper, but with the broader questions of motivation and general framework of analysis. These questions are of the greatest importance, especially in a new field such as nonlinear control system analysis, where a well-chosen initial approach is often the key to success. I would like to summarize in some detail the mathematical and intuitive background from which the paper arose. In so doing, I will express some perhaps intensely personal points of view, which can only be justified, strictly speaking, by their apparent success as far as the present paper is concerned.

In linear control system analysis, the task is clear: (a) Decide whether the system is stable, i.e., whether all the real parts of the roots of the eigenvalue equation are negative; (b) examine in more detail the response to some prototype input, such as a step, calculating initial rise time, settling time, etc.; (c) determine the response of the system to broad classes of signals to which the system will actually be called upon to respond and which are defined perhaps only in statistical terms. These questions are readily agreed upon and interest centers mainly on expedient mathematical methods to answer them.

With nonlinear systems, the situation becomes vastly more complicated. Paralleling the approach to the linear case, the difficulties may be outlined as follows:

1. To start, a way must be found of classifying nonlinear systems from the sta-

bility point of view. A linear time-invariant system is either stable or unstable, regardless of the input signals. In nonlinear systems, stability depends directly on the inputs (or initial conditions) and many phenonema occur: several different types of critical points, limit cycles (steady-state oscillations), regions of completely stable and completely unstable operation, regions in which the character of transient response depends strongly on the amplitude of the input signal, systems which appear docile when subjected to a certain class of signals (steps and ramps) but behave very queerly in response to another class of signals (sinusoids), etc. The list is probably incomplete and much more remains to be discovered.

Clearly then, nonlinear systems should be distinguished according to their peculiarities. If they have no peculiarities, they are not, basically, much different from linear systems and should be studied in much the same way as the latter. The situation is similar, though much more complicated, to that in function theory: Functions are classified according to their singularities; if a function has no singularities, it is very "uninteresting."[1]

To discover these peculiarities, it is essential to be able to visualize the qualitative features of the behavior of a nonlinear system over its entire operating range. The appropriate mathematical tool for this task is the phase-plane method. Dr. Loeb emphasizes, as was insisted on also in an earlier paper,[2] that the phase-plane method is primarily an analytical tool. This basic point has been much obscured in the past since the phase-plane idea is often used for constructing graphical methods of solution. If the phase plane is used for such purposes, it competes merely with the already highly developed techniques of machine computation without offering any significant advantages over the latter except for those who have no access to computers. Only rarely does a graphical method also give insight into a problem. These exceptions are, as far as I know, the Liénard[3] construction and a recent generalization of it by G. Cahen.[4] In these methods, the ideas underlying the graphical constructions also aid significantly in the analysis.

Concentrating on qualitative aspects of nonlinear systems is not only useful but absolutely essential. A well-known mathematical theorem[5] states that the characteristics of a nonlinear system are invariant under any bicontinuous sense-preserving transformation (O-homeomorphism). Intuitively, this means the following: If the trajectories of a system are drawn on a rubber sheet, any deformation (stretching) of the rubber will (nonlinearly) distort the trajectories, thereby leading to a superficially new system, which, however, differs in no essential way from the original one. Thus, for any given nonlinear system there are an infinite number of others which have basically the same properties but are described by different equations, contain different nonlinearities, etc. The theorem certainly agrees with common sense for it would hardly be expected that the dynamic behavior of a control system containing a

saturating amplifier will be altered substantially by replacing the saturating amplifier with another one having a somewhat different saturation curve. For this reason, it is very naive to try to obtain quantitative analytic solutions for most nonlinear problems. The matter of a quantitative solution should and must be left up to a computer or to measurement on a model, and mathematical investigations should be directed toward obtaining general qualitative insight. As a result of these considerations, even crude approximations of nonlinear differential equations will be satisfactory, provided that the approximation retains all the qualitative characteristics of the original system. This idea may be called "approximation in the large."

Discussers have commented on the approximation by straight-line segments. This method stems from the arguments just mentioned, and has been considered in a broader setting in an earlier paper.[2]

Professor Klotter's comment that "increased accuracy entails highly increased labor" is of little significance. The purpose of the straight-line (in-the-large) approximation is to capture in as simple a manner as possible significant aspects of system behavior. Determining the minimum number of segments needed to do this will require further serious mathematical investigation; however, in simple cases, such as the saturation-type nonlinearity (Fig. 3 of the paper), it seems intuitively clear that three segments are adequate. This is not a question of accuracy in the ordinary sense since that problem, as pointed out, should be taken care of by means of a computer.

Professor Ku's remark about the continuity of the trajectories in 3-dimensional phase space should be kept in mind when applying the approximation by dominant terms since otherwise some apparent contradictions may be thought to occur. But in analytical work the precise shape of the trajectories matters little and consideration of the dominant terms in the manner and under the assumptions stated in the paper is quite adequate. For this reason, no check on the accuracy was made, with the exception of the indirect check on the derivation of the quasi-optimum reversal curve in example 3 in the paper, nor is such a check thought to be of particular significance. The conditions under which a dominant-term approximation may be logically used was discussed in some detail in connection with example 4. Again, the guiding principle of the approximation is to retain all essential qualitative features of the original system.

Dr. Loeb's observation that in different regions of the phase space different pairs of principal variables may arise and therefore different normal-co-ordinate transformations are needed, pinpoints a major difficulty, or rather inconvenience, of the analytical procedure as presented. In such a case, e.g., examples 4 and 5 of the paper, it is necessary to resort to much more computation and the interpretation of the approximations may become somewhat delicate. Further improvements can probably be expected as more experience is gained in applying the method to real problems.

Professor Klotter's opinion that "intrinsically, no extension of the phase-plane method to high-order nonlinear differential equations is involved" should not be accepted. The logical criterion by which such a question can be judged is whether the method is capable of explaining problems and phenomena which cannot be predicted or solved on the basis of second-order considerations. Examples 3 and 6 certainly represent inherently higher order systems whose behavior is correctly revealed by the method. The fact that the examples in the paper were all of the third order was intended merely to save algebraic labor.

3. Dr. Weiss brings out many basic and unfortunately unsolved problems. The idea of examining the "latter phases of transient motion," and particularly example 3, was motivated by trying to assess and improve the stability properties of systems subjected to very large signals. By analogy, this means considering the sluing mode of operation of a gun-director (moving from one target to the next), whereas Dr. Weiss's question refers to the tracking mode of operation. These are many practical cases when both modes of operation are equally important.

If the signals in the tracking mode are small, approximately linear operation may be assumed, since any continuous nonlinearity may be considered linear over a small range. This argument merely dodges the problem, although it sometimes closely approximates the practical situation. If the nonlinearity is discontinuous (such as an on-off device), or if the signal, perhaps aided by superposed random noise, makes excursions into the saturation region, the situation becomes very difficult to analyze. Principally, the trouble is that in dealing with a piecewise-linear (in-the-large) approximation of a nonlinear system, it is essential to know when the transient crosses the boundaries of the linear regions. This poses a major problem even when the input is analytically defined (say, a sinusoid). When the input is random the crossing can only be defined in a probabilistic sense, which greatly increases the difficulties. In fact, I am unaware of the existence of a single paper dealing with a probabilistic model of the phase plane.

Much of further success in the nonlinear field depends on vigorous interchange of knowledge about solved or analyzed problems. I would be grateful for any account of instances where the method was applied or where it led to difficulties.

REFERENCES

1. THE MATHEMATICS OF CIRCUIT ANALYSIS (book), E. A. Guillemin. John Wiley and Sons, Inc., New York, N. Y., 1949, p. 266.

2. Reference 13 of the paper.

3. ÉTUDE DES OSCILLATIONS ENTRETENUES, (STUDY OF SELF-EXCITED OSCILLATIONS), A. Liénard. Revue Générale d'Électricité, Paris, France, vol. 23, 1928, pp. 901–12, 946–54.

4. SYSTÈMES ÉLECTROMÉCANIQUES NON LINÉAIRES (NONLINEAR ELECTROMECHANICAL SYSTEMS), A. Cahen. Ibid., vol. 62 1953, pp. 277–93.

5. GLOBAL STRUCTURE OF ORDINARY DIFFERENTIAL EQUATIONS IN THE PLANE, L. Markus. Transactions, American Mathematical Society, New York, N. Y., vol. 76, 1954, pp. 127–48.

Cathodic Protection Circuits

E. W. SCHWARZ
NONMEMBER AIEE

R. M. WAINWRIGHT
MEMBER AIEE

BEFORE discussing the present state of investigation and proposed future considerations of cathodic protection circuitry, it may be profitable to review briefly the problems posed by the observed phenomena of corrosion, in particular the corrosion of metallic surfaces in soils and solutions with which this paper will be primarily concerned. Such a review should be helpful to investigators actively engaged in corrosion studies by emphasizing the actual corrosion problems under consideration as well as present methods of approach. Such a statement of the problem helps in clarifying and unifying much of the apparently disassociated work being done in the particular field and will provide a starting point for investigating corrosion phenomena. Most of the available literature pertaining to corrosion assumes that everyone is quite familiar with the subject and therefore needs no further introduction or explanation. Such an assumption, of course, can frequently be erroneous and, for this reason, it was felt that a brief review of this subject matter would be appropriate.

Statement of the Problem

To give a statement of the corrosion problem, it is necessary to discuss the phenomena that are observed in the physical system. It is found that a metallic surface when placed in a soil, or other chemical environment, acts in two principal ways: the surface becomes anodic—in this case the metallic ions are said to migrate from the surface, combine chemically with ions within the solution, and irreversibly form compounds having no useful metallic characteristics; or, the surface becomes cathodic—in this case it is said that the ionic action prevents metallic ions from leaving the surface although some chemical action may take place, such as production of hydrogen. The anode areas are observed to corrode; the cathode areas are observed

to have no appreciable corrosion in most cases.

Any given small segment of a bigger submerged surface may be either all cathodic or all anodic, or may be divided microscopically into many smaller areas of both classes. The only restriction is that of a whole submerged surface; some of both classes of areas must be present at a given time. A simple example of this type of corrosion is found in the chemical cell of Fig. 1.

As an example of the so-called galvanic cell, electrode 1 may be of copper, electrode 2 of zinc, the solution a wet soil, and the conducting connection may be a wire connecting the copper and zinc; all four of these items must be present in one form or another to have galvanic corrosion at the anode. In many cases items 1 and 2 may be simply different portions of the same metallic area, small differences in metal structure or in solution being the apparent determining factor of corrosion or no corrosion.

The principal physical observations are, then, that the areas identified as anodes corrode, and other cathode areas do not. The economic implication of the anode area action is tremendous. It has been said that the deterioration in underground metal structures caused by galvanic corrosion is in the magnitude of billions of dollars per year in this country. Obviously, the problems of corrosion and corrosion control are of sufficient economic importance to demand that something should be done from the standpoint of engineering. The question of what to do is important.

The present state of technology is such that many different approaches to a problem are available, any one of which may be as fruitful as any other. It is desirable to use the approach having the best chance of solving the problem in the least amount of time and by the least amount of effort. Unfortunately, it is not always possible to determine beforehand which approach will accomplish this solution in the optimum manner. In the corrosion problem, two methods which have been used are the empirical approach and the theoretical approach. Other names might be used for these two methods, but the foregoing names will be used herein.

Briefly, the empirical approach consists of the following procedure: A large

amount of observational field data are taken and recorded pertaining to the corrosion of metallic surfaces. It is then hoped that, by the accumulation of large amounts of these data, empirical relations can be evolved such that corrosion can be predicted, reduced, and perhaps eliminated. This approach has been used in many branches of science and has been quite useful. If the phenomena being investigated are not too complicated, the empirical approach is usually found to be practical and useful. The method does not require a large number of highly trained personnel, and the accumulation of data is generally a quite routine process. It is only the interpretation of the results, for the purpose of deriving empirical relations, that requires highly trained personnel.

Despite the widespread use of the empirical method, there are, nevertheless, certain disadvantages attached to its use. If the problem is relatively complicated, it is found that an abnormally large amount of data must be accumulated before solutions can be obtained at all. Such accumulations of data may lead to large expenditures of time and money. Hence an empirical approach may become undesirable if the problem is highly complicated. Apparently the corrosion problem is complicated to such an extent that the empirical approach gives results of limited, although perhaps useful, levels. The alternative, then, may be the theoretical approach for increased knowledge in the area, as will be now considered.

The theoretical approach consists first of duplicating, in the laboratory, the various conditions existing in the physical system. That is, the corroding system is set up in the laboratory where certain of the variables may be more carefully controlled. The data obtained in the laboratory are analyzed and use is made of existing knowledge in other branches of science. The attempt is made to develop from these laboratory measurements a mathematical formulation, usually in equation form, of the phenomena observed. By mathematical manipulation and extrapolation, predictions may be made involving the laboratory phenomena under varied conditions and, eventually, predictions applicable to the field situations. The end result should be increased knolwedge of the phenomena, explanations, mathematical relations, valid predictions, better measurement techniques, and other improvements in engineering which will be of economic gain in field operations on corrosion control.

This approach, of course, requires the

Fig. 1. The simple galvanic cell

The Equivalent Electric Network

One of the extremely useful techniques used in the theoretical approach to many physical problems has been the equivalent electric network. Since the area of electric circuit analysis is a highly developed area with many useful ways of analyzing and predicting behavior, the electrical analogue or model of the physical system under study gives useful results with a minimum of time and effort. This paper is concerned, therefore, with the implications and use of this equivalent circuit method, and some of the problems and requirements of its use are outlined to increase the knowledge of corrosion and corrosion control.

One of the problems today in the field of transistors is to obtain equivalent electric networks which enable the electrical engineer to apply presently developed analysis methods to the new situation
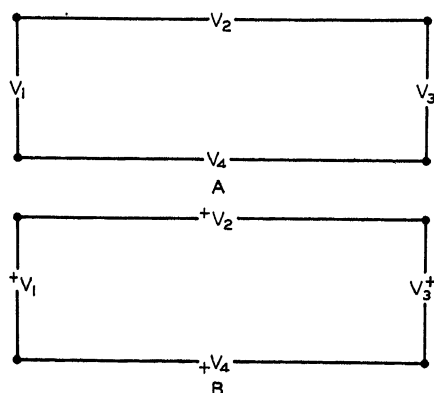
Fig. 2. KVL illustration

involving semiconducting materials. When such a network or series of networks is eventually developed, it must never be assumed that the parts of the network have any exact counterpart among the components of the physical system. In fact, the resistors, capacitors, inductors, switches, source voltages and currents, measured voltages and currents, etc., of the equivalent network should not be assumed to have physical reality at all. The best that can be said is that by the use of these concepts the external characteristics of the physical device have been approximated to some useful end.

One subject frequently misunderstood and overlooked by many corrosion engineers today is the formulation of mathematical equations as derived from the networks. The concept of reference marks for all currents and voltages, for example, is of prime importance. Both of these quantities have sense, or orientation, as well as magnitude. The real number system has the property of giving more information than simple magnitude, i.e., algebraic signs are available to divide the magnitude information into two classes; $+$ and $-$. Thus, if four voltages are arranged circuitally, as shown in Fig. 2(A), without knowing the actual voltages involved, Kirchhoff's voltage law (KVL) cannot be expressed by using the symbols given. However, if reference polarity marks are placed on the diagram, as shown in Fig. 2(B), the KVL equations can readily be stated as

$$-V_1 + V_2 + V_3 - V_4 = 0 \qquad (1)$$

or

$$V_1 - V_2 - V_3 + V_4 = 0 \qquad (2)$$

Either equation 1 or 2 is equally correct here. Now, it may be found that a standardized center-zero voltmeter placed with its $\pm$ terminal at the reference mark will give, for example, numbers as follows

$V_1 = +10$ volts

$V_2 = +15$ volts

$V_3 = -25$ volts

$V_4 = -20$ volts

The $+$ numbers signify, for instance, right-hand deflections of the meter, and $-$ numbers left-hand deflections. KVL is seen to be satisfied by either equation 1 or 2 after placing the numbers in the equations.

A similar situation exists for currents. No equation can be correctly written in symbols only or partly without the reference marking system. In this paper a $+$ mark is used for voltage references and an $\rightarrow$ for current reference. In a double-
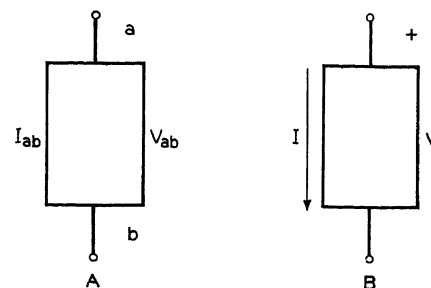
Fig. 3. Current and voltage reference notation

subscript notation, a and b of Fig. 3 are equivalent. Reference marks can be placed on a circuit without any consideration being given for any actual polarities of voltage or actual directions of current already known or surmised. Similar notation should be used in discussing energy transfer and, in fact, any quantity possessing both magnitude and some other sense or orientation property.

After a network has been formed, and mathematical relations written through the use of KVL and Kirchhoff's current law (KCL), analysis can go forward using established techniques of mesh equations, node equations, and various network theorems. Roughly speaking, mesh equations involve writing a series of KVL equations around certain circuit paths, and node equations involve writing a set of KCL equations about $(n-1)$ voltage measuring terminals within the network. Network theorems used may include Thevenin's theorem, Norton's theorem, and so forth. Other techniques such as transformation of variables, Laplace's transforms, nonlinear methods, and the like may be used as needed. The over-all objective will be to make deductions concerning the equivalent network which may be valid in the laboratory model and, consequently, in the original field case.

Duplication of Field Conditions in the Laboratory and Derivation of the Equivalent Electric Network for the Corroding Cell

For the theoretical approach discussed previously, it is necessary to set up in the laboratory the conditions that exist in the physical system under examination.
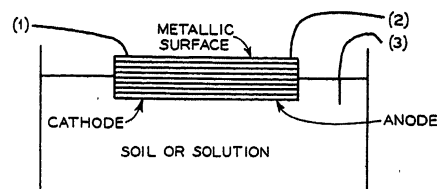
Fig. 4. The laboratory galvanic cell

use of more highly trained personnel than the empirical approach. However, the theoretical approach has more possibilities of solving a complicated problem, as indeed the corrosion problem seems to be, than the former. It has also been found that the theoretical approach will generally provide a solution to a particular problem in a shorter length of time than the more time-consuming empirical methods. This time gain, in fact, forms a major portion of the advantages of the theoretical method.

Fig. 5. The galvanic cell with reference half-cell

Fig. 6. Galvanic cell with third electrode

The easiest way to accomplish this is by placing a metallic surface in a container holding a soil or solution. A simple example is shown in Fig. 4. When a cell is set up in this manner, it is found that under certain conditions the metallic surface can be classified into anode and cathode areas, with corrosion taking place at the former. By the use of this arrangement, the problem of the field has been transferred to the laboratory.

Two questions now arise: Does this laboratory model reproduce with sufficient accuracy the field conditions in which the engineer is interested?; and how can an equivalent electric network be derived?

For the purposes of this paper it will be assumed that the behavior of the laboratory model does correlate with the field observations for the areas of interest—actually this assumption would have to be proved more conclusively at some later date—and the question of how the network is to be constructed follows immediately.

Electric network theory assumes that voltage and current measurements and energy transfers are an inherent part of the analysis picture, and that current and voltage behavior patterns associated with a 2-terminal network are important. This leads to a necessity for making current and voltage measurements on the laboratory model or finding other types of measurements which are analogous to these. Fortunately for the galvanic cell of Fig. 4, certain electrical measurements are possible, but not without some modification of the cell. For example, voltages could be measured between points 1 and 2 of Fig. 4, but, since these points are found in practice to have extremely small voltage values, such measurements are not too useful. Measurements of voltage between points 2 and 3 are possible, but only with some change in the cell. Since point 3 represents merely a point in the solution, it is not available as a conducting terminal for a meter. Any strip of

metal introduced at point 3 as a terminal, develops itself voltages which are extraneous to the desired voltages.

There is no immediate answer to this problem of finding a suitable terminal for the meter but, fortunately, a permanent and useful modification of the original cell can be made which does not appear to destroy the validity of the original cell. This modification takes the form of a physical arrangement called a half-cell, as shown in Fig. 5.

This half-cell consists of solution Y in contact with the original soil or solution X. Metal N then contacts the new solution which is prevented from diffusing into X by a porous membrane or plug. The total voltage measured is, then, $V_{23} + V_{34} = V_{24}$. The advantage of this scheme is that V_{34} is a relatively constant or known voltage and, therefore, V_{24} varies in a known relationship to V_{23}, which condition is sufficient for analysis purposes. The modified cell of Fig. 5 can now be thought of as a 2-terminal network between terminals 2 and 4. Half-cells for this purpose include the saturated calomel cell, usually used in the laboratory, and the copper-copper sulphate cell usually used in the field.

Suppose now that the situation portrayed in Fig. 5 is changed to include a third electrode having points of interest at 5 and 6. Point 2 has been eliminated for simplification here, and the modified cell is shown in Fig. 6. Assume that E_{cp} is a source voltage which can be varied at will. Relationships involving V_m, I_{cp}, and E_{cp} could be determined for example. These relationships and other more complicated experimental work[1,2] done by various scientists suggest that the corroding cell can be represented by the following circuit map of Fig. 7. Here R_1 and R_2 represent resistances within the piece of corroding metal, and may be considered approximately zero. The voltages within the left-hand dotted enclosure represent the surface film voltages for the cathode area, and the voltages in the right-hand enclosure are those for the anode area. R_c and R_a represent solution or soil resistances. The current associated with the corroding circuit is called $i_g(t)$, and it is seen that $i_g(t) = i_c(t) = i_a(t)$ for the references given. Based on previous work[1-4] the surface film voltages are considered representable by a constant source voltage in series with a source voltage ϕ which depends upon current density and time.

The addition of a voltmeter to the foregoing circuit of Fig. 7 gives an added complication, as shown in Fig. 8. The circuit elements for representing the half-cell
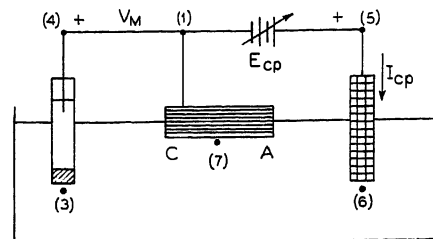
have been constructed[3] based on certain experimental work, and deductions therefrom, and are shown in the figure by the elements within the dotted enclosure of Fig. 8. The voltmeter, reading V_m, of Fig. 8 obviously gives a voltage which varies as V_{17}, even though the meter does not read V_{17} directly.

Examining now the diagram of Fig. 7 or Fig. 8, it could be assumed that corrosion would cease if $i_a(t)$ were reduced to zero. This assumption is based on the presumption that $i_a(t)$ is analogous to the anode current in the laboratory cell. If no metallic ions leave the anode, presumably there will be no current and, hence, no corrosion.

The next question is: How should $i_a(t)$ be reduced to zero in the equivalent network? By previous experience with such networks, one possibility would be the introduction of a current $i_p(t)$, as shown in the the diagram of Fig. 9. E_p is the source of this current, therefore, $i_p(t)$ will correspond to what it usually termed the cathodic protection current.

The complete equivalent network is now available for the corroding and cathodically protected cell, with voltage-measuring devices added, and the situation is now ready for an application of circuit analysis methods.

Analysis of the Equivalent Network

As previously stated, two general methods of analyzing the behavior of a network such as is shown in Fig. 9 are available in the mesh and node equation methods. For the type of circuit being considered here, both methods may be equally applicable.

An application of the node method with one equation written at terminal 1 with terminal 7 as reference node, results in the following

$$i_p(t) + V_{17}(1/R_a + 1/R_c) - E_a + \phi_a/R_a - (E_c + \phi_c)/R_c = 0 \quad (3)$$

This equation assumes that R_1 and R_2 resistances are negligibly small, and that i_m is approximately zero since $R_m + R_{vm} >> R_1 + R_c$.
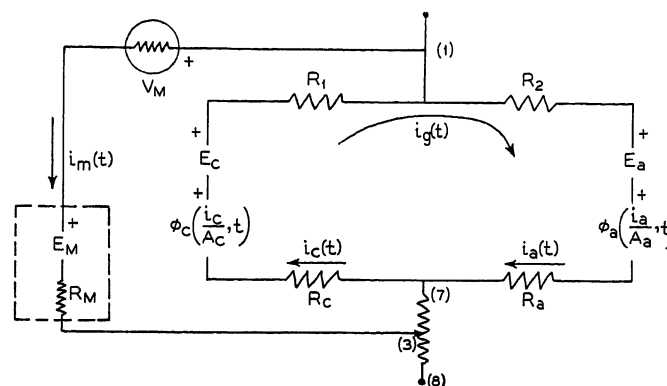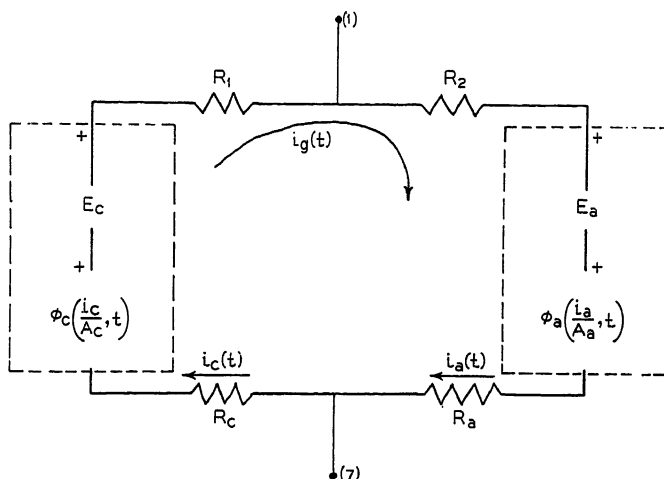
Fig. 7 (left). The equivalent circuit for the corroding cell

Fig. 8 (above). Equivalent circuit for the corroding cell with reference half-cell

The next step is to write a mesh equation involving the anode branch of the circuit and V_{17} as follows

$$V_{17}=E_a+\phi_a+i_a(t)R_a \qquad (4)$$

Now, if $\phi_a=0$, when $i_a(t)=0$, from previous writings,[1-3] then for $i_a(t)=0$

$$V_{17}=E_a \qquad (5)$$

This equation 5 is then a criterion for cathodic protection, under the simplifying assumptions as stated. By substituting the results of equations 4 and 5 back into equation 3, it will be found that

$$i_p(t)=[(E_c-E_a)+\phi_c]/R_c \qquad (6)$$

Equation 6 gives the value of cathodic protection current for no corrosion in the previously anodic branch. The voltage (E_c-E_a) will be recognized as the open-circuit driving voltage of the corroding cell, and ϕ_c as the cathode polarization function. Note again here that the pluses and arrows forming the reference system of marks are not to be interpreted as having significance in determining actual polarities and directions until numbers are used in the equations.

The voltmeter, whose reading V_m is proportional to V_{13}, still does not read V_{17} by the difference voltage V_{37}. A KVL equation involving the voltmeter can be written as

$$V_m=E_a+\phi_a+i_a(t)R_a-E_m-$$
$$i_p(t)R_{37}-i_m(t)R_m \qquad (7)$$

Here, if R_1 and R_2 are considered zero as before, $i_m(t)$ is very small, and $i_p(t)=0$, as in the freely corroding case, this equation reduces to

$$V_m=E_a+\phi_a+i_a(t)R_a-E_m \qquad (8)$$

After cathodic protection is applied with $i_p(t)$ as in equation 6, the voltmeter now reads

$$V_{mc}=E_a-i_p(t)R_{37}-E_m \qquad (9)$$

The net change in voltage from no protection to full protection is now

$$\Delta V_m=V_m-V_{mc}=i_p(t)R_{37}+\phi_a+i_a(t)R_a \qquad (10)$$

If the term $i_p(t)R_{37}$ is made as small as possible and R_a is reduced, corresponding in the physical case to moving the reference half-cell closer to the corroding metal, the change in voltage is approximately

$$\Delta V_m=\phi_a+i_a(t)R_a \qquad (11)$$

An immediate conclusion of this analysis is that the anode polarization function may be of considerable importance in determining the change in potential-to-soil reading which should accompany complete cathodic protection in the field. Further deductions along this line will be left to the interested reader.

The analysis as shown verifies that

corrosion should be substantially reduced to zero if the voltage V_{17} is made equal to the greatest anode open-circuit potential. If there should be many such circuits in parallel, representing more closely the field case, the reduction in corrosion as measured by a change in V_{17} should be a statistical problem based on a distribution of values for E_a as derived from field and laboratory experiments.

Conclusion

A statement of the corrosion problem has been presented along with the present methods of attack—empirical and theoretical. The advantages and disadvantages of these methods have been briefly discussed. The purpose and implications of the equivalent electric network as used in the theoretical approach were shown, in addition to a discussion of how such an
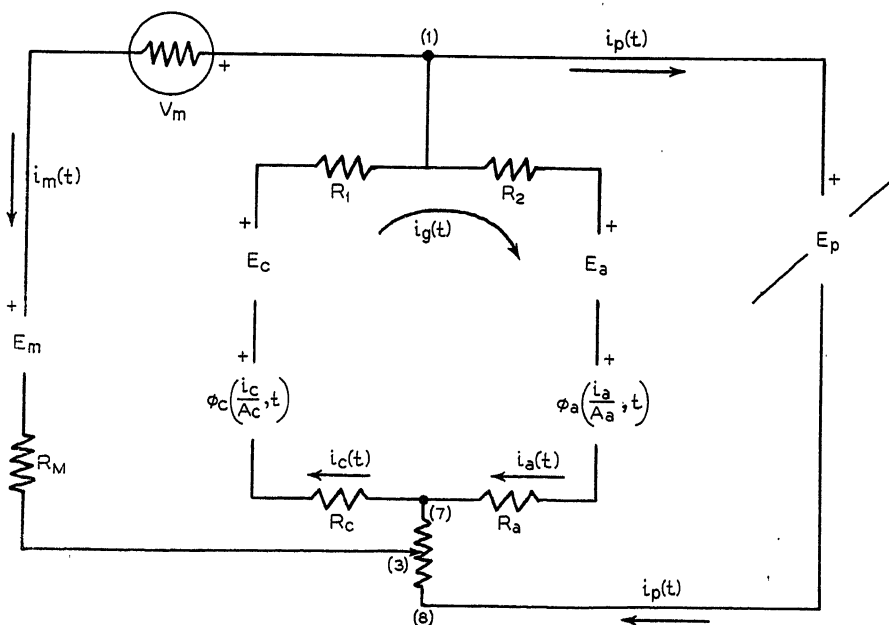
Fig. 9. Equivalent circuit for the corroding cell with reference half-cell and cathodic protection

equivalent network may be evolved for the corroding cell. Once having determined the network, the present-day methods of circuit analysis were employed to establish the criterion for cathodic protection. One important argument for the use of such a theoretical approach results from the fact that the criterion for cathodic protection could have been deduced

directly from consideration of the equivalent network without resort to further field research. It is likely that the problem of the so-called remote-earth reference could also be considered in quite an analogous manner by the use of a multimesh planar resistance network which will approximate the field distribution of currents and voltages.

References

1. R. B. Mears. *Journal and Transactions,* Electrochemical Society, Baltimore, Md., vol. 95, no. 1, 1949.

2. R. B. Mears, Brown. *Journal, Ibid.,* vol. 97, no. 3, 1950, pp. 75–82.

3. Nehama, R. M. Wainwright. *Report No. 3,* Cathodic Protection Laboratory, University of Illinois, Urbana, Ill., June 1, 1953.

4. ALTERNATING CURRENT CIRCUIT THEORY (book), M. B. Reed. Harper and Brothers, New York, N. Y., 1948.

System Neutral Grounding for Chemical Plant Power Systems

D. S. BRERETON
ASSOCIATE MEMBER AIEE

H. N. HICKOK
ASSOCIATE MEMBER AIEE

THE chemical industry, because of its continuous processes, has always demanded reliable performance from its electrical power system. Observation of electrical practices will show that care has been exercised in the selection and application of electrical equipment. The scope of this paper is to review the influence which system neutral grounding has on the performance of electrical apparatus, something of the history of grounded and ungrounded systems and why they have been selected, and a discussion of the methods of system neutral grounding with suggestions on how the over-all performance of the power system and the connected equipment can be improved by the operation of some form of system neutral grounding.

History of Industrial Grounding

System neutral grounding has always been applied to some voltage level in the industrial power field. In the low-voltage class, the 208-volt (or better known as the 120/208 Y-volt) system has, for all practical purposes, been exclusively oper-

ated as a grounded system. For the medium-voltage class, i.e., from 601 to 15,000 volts, the voltage levels of 4,160, and 13,800 volts have predominantly been operated with the system neutral grounding. The voltage levels of 480 volts and 2,400 volts have experienced a considerable degree of ungrounded operation in the past. The last 10 years have seen a great increase in the application of system grounding at the 480-volt level with a decidedly noticeable increase at 2,400 volts in the last 3 years. In retrospect, the most frequent reason given for the ungrounded operation of these two voltage levels has been the claim for greater service reliability because there will not be a tripout for a single line-to-ground fault. Years of experience in many industrial fields have shown that ungrounded systems are not as reliable as grounded systems. Such an example is the experience cited by Arberry,[1] After stating the importance of continuous service for the glass industry, he notes, "With continuous process operations the hunting of ground faults is very difficult, and two grounds on the same phase but on two different feeders are exceedingly difficult to trace. This is because all the feeders must be opened at once and closed one at a time to find the trouble. Our experience is that the first ground fault remains on the system because we cannot open the feeder breakers to hunt it. The result is that the system operates with two phases at line-to-line voltage-to-ground and the operating electrician hopes that no other ground occurs before he has the opportunity to find the first one. It was because of our experience such as I have mentioned, and the

need in our operations for the highest possible service continuity, that we began to seriously consider the use of grounded neutral low-voltage distribution systems."

Arberry's paper continues with a discussion of how he applied system neutral grounding and concludes with the following paragraph: "Our experience with these [neutral grounded] systems has been very satisfactory. There is no question that the service reliability has greatly improved. The majority of the faults occur on branch feeders and are cleared by the local branch protection devices such as fuses. Troubles are localized and promptly repaired. As the electricians become used to the new systems they are more enthusiastic and quickly learn, for instance, that a single blown fuse promptly indicates a ground. None of them have expressed any desire to return to non-grounded systems."

It is the authors' opinion that the most complete and useful information dealing with the causes and the results of abnormal overvoltages in industrial systems has been given by R. H. Kaufmann.[2] Kaufmann's 1952 paper conveniently summarized the various causes of over-voltages and pointed out actual case histories of damage to a power system from the causes given. The overvoltages discussed in the paper were: 1. lightning, 2. static, 3. physical contact with a higher voltage system, 4. resonance effects in series inductive-capacitive circuits, 5. repetitive restrike (intermittent grounds), 6. switching surges, 7. forced current zero interrupting, and 8. autotransformer connections. Kaufmann states, when discussing repetitive restrike on low-voltage systems, "Intermittent ground fault conditions on low-voltage ungrounded neutral systems have been observed to create overvoltages of five or six times normal quite commonly. An unusual case involved a 480-volt ungrounded system. Line-to-ground potentials in excess of 1,200 volts were measured on a test volt meter. The source of trouble was finally traced to an inter-

Fig. 1 (above). System-neutral circuits and methods of grounding

X_G—reactance of generator or transformer used for grounding
X_N—reactance of grounding reactor
R_N—resistance of grounding resistor

Fig. 2 (above right). Low-voltage solid-grounding and medium-voltage solid-grounding for a small system

Fig. 3 (right). Double line-to-ground fault on ungrounded system results in outages of two circuits and high-level fault currents which can cause severe damage to equipment
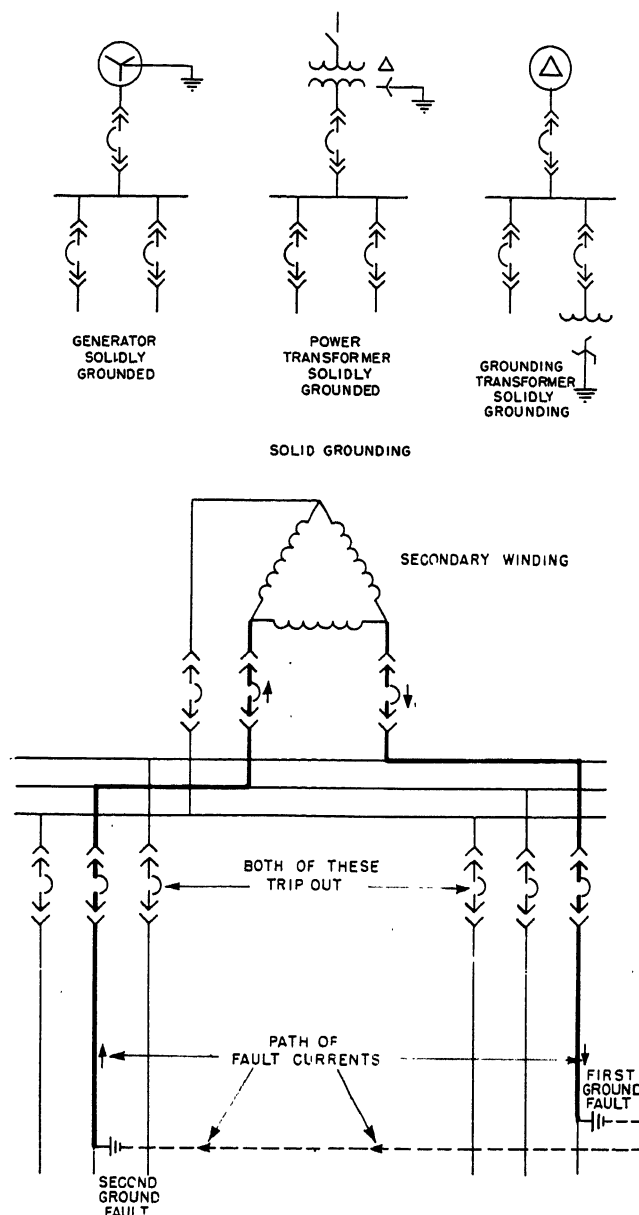
mittent ground fault in a motor starting autotransformer. About two hours lapsed while the source was being located during which time between 40 and 50 motors broke down."

An excellent comparison of a power system, part of which operated ungrounded and the other part grounded, is included in a 1953 AIEE conference paper by C. L. Eichenberg.

In his comparison between the ungrounded and grounded sections of his 6,900-volt system Eichenberg says, "The operating record of the system since the grounded neutral was installed is most gratifying. The ground faults experienced show a marked reduction in number and severity. For instance, during the year 1944 the number of ground indications recorded totaled 34. Of these 34 indications, 19 resulted in equipment

failures such as grounded motor coils or flashed-over bushings. During the year 1951 (with the system grounded) there were two ground relay operations resulting in one equipment failure, and the first fifty weeks of 1952 show a similar record. Particular attention has been paid to the severity of the damage caused by these ground faults. In each instance it appears that the relaying has been fast enough to clear the fault before any destructive burning resulted."

These examples show that it has taken actual field experience to indicate that improved system continuity and reliability can be obtained from the grounded over the ungrounded system. This is very important to the chemical industry since such a substantial percentage of the utilization equipment is now supplied by 480- and 2,400-volt ungrounded systems.

Description of Various Grounding Methods

Many different means have been employed in the chemical and other industries to ground the system neutral. The various methods of grounding, as well as ungrounded operation, are shown in Fig. 1. The general trend in industry is toward solid grounding for low-voltage systems and resistance grounding for medium-voltage systems.

Ungrounded Operation

The operation of the system with the neutral ungrounded, Fig. 1(A), has been proposed as a method of maintaining service to essential loads on the occurrence of the first line-to-ground fault. This desire for not tripping on the occurrence of the first line-to-ground fault

TABLE I. SYSTEM CHARACTERISTICS WITH VARIOUS GROUNDING METHODS

	Ungrounded	High Resistance Grounding	Low Resistance Grounding	Solid Grounding
Properties during faults				
Current for a phase-to-ground fault in percent of 3 phase fault current	Less than 0.1%		5% to 20%	About 100%
Transient overvoltages	Up to 6 times	Not more than 1-1/2 to 2 times normal		
Automatic segregation of faulty circuit and equipment	No		Yes	
	No outage		Outage	
Circuit outage for various types of system faults — One ph.to ground	No outage			Outage
Phase-to-phase) Two ph.to ground) Three phase)	Outage			
Multiple Faults	Many case studies reported showing multiple failures	Report of insignificant number of failures showing multiple outages.		
Power System Properties				
Transformer; winding connection	Delta		Wye or delta with grounding transformer	
Fault Location Method	Have to take part or all of the system out of service or use ground fault locator to find ground faults		System does not have to be taken out of service because the faulty equipment has been automatically isolated	
	If ground fault is not removed, may lose two circuits due to another ground fault		Ground faults are localized and trip off immediately	
First Cost — Low voltage systems	Delta Connected Sub-station with ground detector generally costs more than wye	Slightly higher due to high resistance resistor and ground indicator	Not generally applied, but would be slightly higher	Lowest, in that wye and delta transformers cost about the same
First Cost — Medium voltage system	Including ground detector equipment delta is slightly lower in cost	Somewhat higher due to high resistance grounding equipment	Somewhat higher due to low resistance grounding equipment	Wye connected substation slightly higher than delta
Maintenance cost	Takes time and equipment to find grounds		Ground faults are easily located	
Rating of lightning arresters	Ungrounded neutral type			Grounded-neutral type
Application of grounding method	Less and less frequently applied	Applied on low or medium voltage systems when system not permitted to be tripped for first ground fault	Applied on medium-voltage systems i.e., 2.4, 4.16, 6.9, or 13.8 KV	Applied on low-voltage systems i.e., 208, 240, 480, or 600-Y Some application on small medium voltage systems

has resulted in a wide variety of ground-fault locating equipment to locate more freely and to remove the fault from the system. Experience has indicated that ungrounded operation permits abnormally high transient overvoltages which may cause damage to the connected electrical apparatus.

SOLID GROUNDING

Low-voltage services, 600 volts and below, are more and more frequently solidly grounded; see Fig. 1(B). Such services are generally supplied from load-center unit substations, of which the circuit breakers employ direct-acting trip devices. Solid grounding provides approximately the same amount of ground fault current for a 3-phase fault or a line-to-ground fault; thus, the phase-connected protective devices can provide proper protection. The medium-voltage system might be solidly grounded if the line-to-ground fault current is fairly low, in the order of 3,000 amperes. Methods of solid grounding are shown in Fig. 2.

REACTANCE GROUNDING

The most frequent application of re-actance grounding, Fig. 1(D), has been for low-voltage generators. A generator is usually braced to withstand only its 3-phase fault current. Generally its zero-phase sequence reactance will be less than the positive- or negative-phase sequence reactances resulting in a line-to-ground fault current greater than the 3-phase fault current if the generator is solidly grounded. A reactor may be employed in the system neutral to provide sufficient reactance to limit the line-to-ground fault current to the 3-phase fault current. Generally speaking, if a small impedance is desired in the neutral, a reactor will be found most economical and if a large impedance is required, a resistance will be used.

GROUND-FAULT NEUTRALIZER

The ground-fault neutralizer, Fig. 1(E), has been applied in a very limited number of cases; principally it has been in systems where the plant operator does not desire to trip a circuit on the occurrence of the first ground fault. Another application has been where a degree of ground-fault protection was desired in

large existing systems having only two current transformers per circuit, making the application of the residually connected ground relay difficult. The ground-fault neutralizer has also been applied where the zero-phase sequence charging current has been high.

RESISTANCE GROUNDING

The general trend in industries is for all of the voltage classes of medium-voltage systems, such as 2.4, 4.16, 6.9, and 13.8-kv, to apply a resistance grounded system, Fig. 1(C). These systems usually incorporate power circuit breakers and relays, and can include a residually connected ground relay which can provide much faster and more sensitive protection than a phase-connected device for a line-to-ground fault. This is particularly important in the medium-voltage system because it will have a higher level of 3-phase fault kva as compared to the low-voltage system. It is important to realize that the additional equipment used to ground a power system is a very small percentage of the electrical system cost. Further, the least

expensive way of substantially limiting line-to-ground fault current is by use of a resistor.

Reference has been made to resistance grounding for the industrial system. Nearly all of the published information[3] dealing with this form of system neutral grounding will apply a resistor to limit the line-to-ground fault current from 5 to 20 per cent (%) of the 3-phase fault current. This type of system will be so arranged that the ground relays will trip the circuit if a fault should occur. This type of resistance grounding will now be referred to as "low-resistance grounding." A recent type of resistance grounding, called "high-resistance grounding" will apply to the system when the largest practical value of resistance is applied in the system neutral. This will result in a line-to-ground fault current of less than 0.1% of the 3-phase fault current. No means are provided for removing a faulted circuit for a single line-to-ground fault.

HIGH-RESISTANCE GROUNDING

High-resistance grounding, defined as a resistance nearly equal to, but not greater than, the zero-phase sequence capacitive reactance of a system, $1/3\ X_{c_0}$, may be applied in the chemical or any other industry where it is desired to limit transient overvoltages and not trip out a circuit or a piece of equipment upon the occurrence of a single line-to-ground fault. The range of zero-phase sequence reactance for the typical chemical industry power system may be from 50 to 500 ohms for a 480-volt system and from 500 to 10,000 ohms for a 2,400-volt system. Since the high-resistance grounded system does not have all the properties of the low-resistance grounded system, such as being able automatically to isolate faulty equipment, it is desirable to state the characteristics of this additional form of system-neutral grounding. Simply stated this means that continuous process industries, such as the chemical industry, may propose that an attempt be made to keep a faulted circuit in service, but that such a fault should not cause damage to the other connected apparatus or cause serious overvoltages that may cause multiple circuit failures. Kaufmann, previously mentioned, explains the many possible ways that transient overvoltages may occur on the ungrounded system. As evidence is kept and accumulated on the operation of ungrounded systems, it soon becomes apparent that multiple failures frequently occur on ungrounded systems. It has been shown that transient overvoltages, in the neighborhood of six times normal, can be sustained on the

ungrounded system. This stress is imposed on all the insulation of all motors, cable, and other electrical apparatus connected to the same metallic circuit for the duration of the fault. High-resistance grounding moderates to ineffective values these transient overvoltages for a single line-to-ground fault.

Comparison of System Properties During Faults for Various Grounding Methods

A comparison of system characteristics with various grounding methods is given in Table I. The first group of data in this table deals with the properties and performance of the system during a fault. It should be noted that although transient overvoltages are eliminated with high-resistance grounding, as against an ungrounded system, neither of these systems provides automatic segregation of the faulty circuit and equipment. Automatic segregation would occur for two feeder services in the case of a second line-to-ground fault as shown in Fig. 3. Experience has indicated that the most frequent type of fault occurring in the industrial system is the line-to-ground fault. Neither the high-resistance grounded nor ungrounded system will remove the first faulted circuit, should a fault occur in an electrical equipment. Should any other type of fault occur, such as a phase-to-phase or 2-phase-to-ground or a 3-phase fault, the faulty equipment will automatically be isolated.

Comparison of Power System Properties for Various Grounding Methods

Under the section in Table I dealing with power system properties, a comparison is given of the various fault location methods for the four types of system operation. Low-resistance grounding or solid grounding automatically isolates the faulted equipment. Determining the circuit on which the fault occurs can be very time-consuming and costly when some methods of operation are used. This is particularly true if a fault should occur on the same phase of two different services for that voltage level.

For low-voltage systems, the comparison of first cost shows that the solidly grounded neutral is lower because the delta- and wye-connected transformers are nearly the same price and the delta-connected transformer requires the additional expense of ground indicator equipment to tell when a ground fault exists on the system. High-resistance grounding,

with a ground resistor and a ground indicating relay, may cost a few dollars more than an ungrounded system with ground indicators; see Fig. 4. In comparing the various methods of grounding for a medium-voltage system, the ungrounded, or delta-connected system may be slightly lower in cost because of the fourth bushing required for the wye-connected transformer. When high-resistance grounding is applied on the medium-voltage system, two methods should be checked to determine the lowest first cost. A medium-voltage high-resistance resistor may be inserted directly in the neutral of the wye-connected transformer, Fig. 5, or, and this may be more economical, a small distribution transformer may be applied in the neutral circuit and loaded with a resistor to provide an equivalent resistance in the neutral circuit; see Fig. 4. In the case of both low- and medium-voltage existing ungrounded systems high-resistance grounding may be economically applied by the use of a 3-phase 2-winding transformer or three single-phase 2-winding transformers. This is shown in Fig. 6. Indication of a line-to-ground fault is provided by the voltage relay applied across the resistor. Protection of the transformers can conveniently be provided by the delta-connected current transformers. This current transformer connection will circulate zero-phase sequence currents, as appearing during ground faults, and permit the detection of positive- or negative-phase sequence components of current by the time-over-current relays. It should be emphasized that the slight additional cost for equipment used in providing system neutral grounding is a very small portion of the total cost of the electrical system and is small in proportion to the benefits of added service reliability obtained from it.

An important comparison will be that of maintenance costs. The high-resistance grounded system has the very great advantage over an ungrounded system of reducing transient overvoltages, thus lengthening the life of all the apparatus connected at that voltage level, but still requiring the same time and equipment to find ground faults as does the ungrounded system. The elimination of multiple failures caused by transient overvoltages inherently means that less apparatus is damaged, resulting in much lower maintenance and replacement costs.

Application of High-Resistance Grounding Method

One of the most important comparisons in Table I deals with the application of
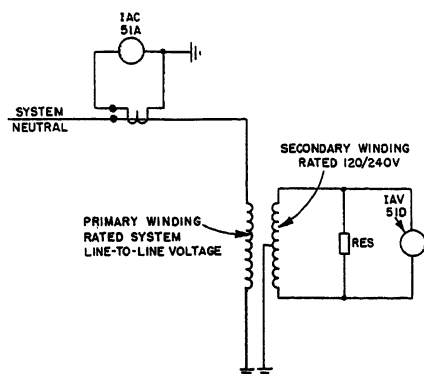
Fig. 4 (left). Use of 2-winding transformer for medium-voltage high-resistance grounding when system neutral is available
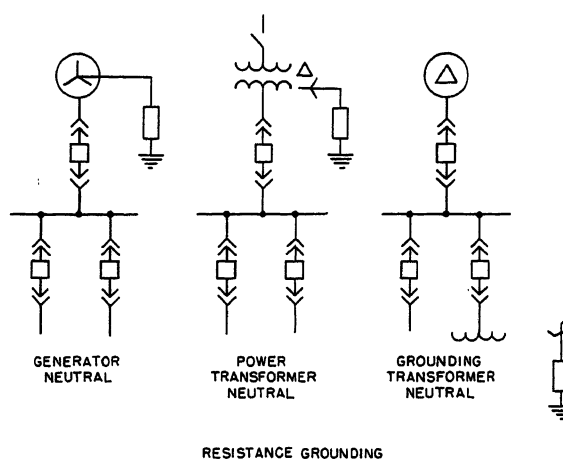
Fig. 5 (right). Low-voltage high-resistance grounding and medium-voltage low- and high-resistance grounding

GENERATOR NEUTRAL   POWER TRANSFORMER NEUTRAL   GROUNDING TRANSFORMER NEUTRAL

RESISTANCE GROUNDING

grounding methods. There is every indication that ungrounded operation is rapidly declining in usage. When a portion of an operation in a chemical plant, or other such industry, will not permit tripping on the first ground fault, a high-resistance grounded system may be employed. In general, the preferred application is for low-resistance grounding on medium-voltage systems and solid grounding on low-voltage systems to isolate a fault automatically, which saves hunting for it.

LINE-TO-GROUND CURRENT FOR HIGH-RESISTANCE GROUNDING

Investigation has shown that there is a very limited amount of information dealing with the zero-phase sequence charging current of motors[4]. A good deal of information[5] is available on the charging current for transformers, autotransformers, high-voltage potential transformers, high-voltage current transformers, induction regulators, current-limiting reactors, power circuit breakers and bushings, insulators, and lightning arresters, as well as other apparatus, because of the extensive amount of work

Fig. 6. Use of 2-winding transformer(s) for low- or medium-voltage high-resistance grounding when system neutral is not available

done in the investigation of recovery voltages. Data are currently being accumulated to determine parameters dealing with the charging current for motors. This is particularly important in the application of high-resistance grounding because motors are the major contributor of charging current in the modern industrial power system.

Table II, showing the results of the investigation of the zero-phase sequence capacitive property of the industrial system, may be used as a general guide for high-resistance grounding. It deals with the application of high-resistance grounding, and is based on the operation of a load-center type of distribution system with approximately 100% connected motor load. An extensive amount of cable would decrease the zero-phase sequence capacitive reactance (increasing the charging current) and a lower percentage of motors would increase the capacitive reactance (decreasing the charging current). The value of the grounding resistor suggested in the table permits some allowance for system expansion, but does not provide for the application of surge capacitor equipment to motors. Where rotating machine protective capacitors are applied, it will be sufficiently accurate to add this additional capacitive current to the current given in Table II to determine the lower ohmic rating of the system neutral grounding resistor.

PROTECTION OF EQUIPMENT FOR HIGH-RESISTANCE GROUNDING

The method of detecting when a ground fault occurs in a system will be to observe the zero-phase sequence voltage across the high-resistance resistor in the system neutral or, as may be applied in the medium-voltage system, across the loading resistor on the secondary of the single-phase distribution transformer. An induction-type voltage relay, provided

with third-harmonic compensation, may be applied, such as the General Electric IAV51D. This relay is designed so that its third-harmonic voltage pickup is approximately three times greater than its fundamental zero-phase sequence voltage pickup. An overcurrent relay could also be applied, but it is not readily available with third-harmonic compensation.

The high-resistance system does not permit sufficient ground-fault current to trip protective devices. To prevent the ground fault from being maintained on the system for an extensive period, a timing relay might be operated from the voltage relay, as well as an alarm. One plant at present applies a timing relay set for a maximum of 2 hours to permit other equipment to be substituted for the faulty equipment before the timing relay initiates the tripping circuit.

A chemical plant operating a 2,400-volt ungrounded system may choose to provide the neutral by use of a zigzag grounding transformer; see Fig. 5. As has been stated, either a medium-voltage high-resistance resistor may be applied,
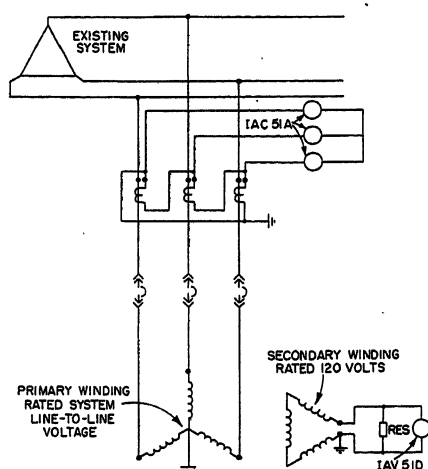
Table II. Application of High-Resistance Grounding

System Voltage Volts	Connected Load at Voltage Level Kva	System Neutral Grounding Resistor Ohms	Line-to-ground Fault Current Amperes
480 ..	{1,000 kva or under....	90......	3
	{1,500 to 3,000 kva....	45......	6
2400*..	{2,500 kva or under....	280......	5
	{3,750 to 7,000 kva....	140......	10

*The standard rotating machine protective capacitor is rated 0.5 microfarad per pole. For a line-to-ground fault, a set of capacitors would contribute 0.785 ampere to the capacitive charging current.

$$3Ic_0 = (V_{L-n})(2\pi fC) = (1388)(377)(3 \times 0.5)/10^6$$
$$= 0.785 \text{ ampere}$$

If ten motors on the 2,400-volt system had protective capacitors, the resistor size should be decreased to permit an increased current of approximately 8 amperes.

or a single-phase distribution transformer with a loading resistor to provide an equivalent resistance in the system neutral; see Fig. 4. Normally, no voltage (with the possible exception of a small value of third-harmonic voltage) would exist across the secondary of the distribution transformer. Should the single-phase secondary of this transformer become short-circuited, there is no convenient means of detection prior to a ground fault appearing in the power system. To assure proper protection under such circumstances, it is suggested that a current transformer be placed between the distribution transformer or high-resistance resistor to operate an overcurrent relay (shown as the *IAC51A* in Fig. 4) to provide protection against a short circuit of the secondary or primary of the distribution transformer. The current transformer in this circuit should have a mechanical limit greater than the maximum

symmetrical line-to-ground fault current.

Conclusions

The chemical industry has contributed much to the electrical art. As this industry expands and applies newer methods, it is hoped that this review of system neutral grounding will not only provide a review but present suggestions that will permit improved performance of the electrical power system through the grounding of power systems at all voltage levels. The grounding method should be selected that best suits the process and operating conditions. When tripouts are permitted, a solidly grounded or low-resistance grounding system should be applied. When tripouts are not desired, the high-resistance grounded system may be applied. The increased

availability of ground-fault detection equipment, which more easily permits the location of a round fault, makes practical the high-resistance grounded system.

References

1. AIEE Power Conference: Power Generation and Industrial Power Systems. *AIEE Special Publication S-38*, "The Use of 600 Volt and 460 Volt Power Systems with Grounded Neutrals," James P. E. Arberry. Aug. 1950, pp. 187–88.

2. Neutral Grounding of Low-Voltage Systems, R. H. Kaufmann. *Iron and Steel Engineer*, Pittsburgh, Pa., Feb. 1952, pp. 96–103.

3. Industrial Power Systems Handbook, edited by D. L. Beeman. McGraw-Hill Book Company, New York, N. Y., 1955.

4. Capacitance of Synchronous-Machine Armature Windings Determined for High-Potential Test, R. W. Wieseman. *General Electric Review*, Schenectady, N. Y., July 1947, pp. 26–30.

5. Power System Overvoltages Produced by Faults and Switching Operation, AIEE Committee Report. *AIEE Transactions*, vol. 67, pt. II, 1948, pp. 912–22.

Automation for Gravity Freight Classification Yards

A. V. DASBURG
MEMBER AIEE

Synopsis: Recently developed speed-measuring and control systems automatically retard cars and guide them to classification tracks in modern gravity yards. Uncoupled from the train at the crest, "cuts" consisting of one car, or several coupled cars, having wide variations in rolling resistance select prior established routes and determine their own releasing speeds as they move by gravity to couple safely with other cars in the yard.

THE simplest form of freight car classification yard is the original form of a group of parallel tracks, connected on both ends, on which a locomotive shuttles back and forth to sort cars into proper order before making them up into trains. Bringing mechanization and automation to this basic operation has been a step-by-step process.

Increases in freight traffic required faster sorting of cars, and in 1883 the first step was taken by placing one end of the yard on a rise or "hump." As a locomotive steadily pushed cars to the crest they were uncoupled and ran by gravity to the classification tracks. Brakemen rode the cars to prevent them

from coupling at damaging speeds in the classification tracks. Other men were assigned the duty of operating switches along routes followed by the cars.

Power-operated switches were introduced in 1891, permitting one man to control several switches. In 1924, car retarders (sometimes called rail brakes) were perfected and these have largely replaced the car riders.

A simple form of gravity yard designed for retarder operation is shown in Fig. 1, and its profile diagram in Fig. 2. The scale permits automatic weighing of cars as they roll down the grade. Gravity operation is now used extensively for major yards, but flat yards continue to be used for small-volume sorting operations.

Automatic control of retarders to provide a fixed releasing speed was introduced in 1941 and expanded in 1951 to provide multiple speed selection. This scheme uses a series of short electric track circuits. When shunted by the wheels they detect the position of a car in the retarder. Speed is determined by timing the intervals between shuntings. Control

of retarder pressure is derived from this and a release speed setting chosen by the retarder operator.

The development of automatic route switching in 1950 relieved the retarder operator of switch operation and permitted control, by one operator, of a yard with any practical number of tracks. Routes are established on a relay network by pressing destination buttons on a route selection machine (Fig. 3) located at the crest of the hump. As the cuts roll forward they are detected by track circuits associated with each switch and automatically call for routes corresponding to the destinations selected. Route storage for an entire train can be added to this system when circumstances warrant. Usually the route machine is controlled by a member of the engine crew at the hump. Thus the retarder operator has only the task of adjusting retarder pressure so as to maintain enough separation between cars to permit switch operation, and to release cars at speeds required for safe coupling on the classification tracks.

Solutions under manual operation are developed intuitively by an operator from the experience of handling thousands of cars of various types under a variety

D | ZONE I | ZONE 2

SCALE
CREST
HUMP
RETARDER
A
B
C
GROUP
RETARDERS
CLASSIFICATION TRACKS
(NOT TO SCALE)
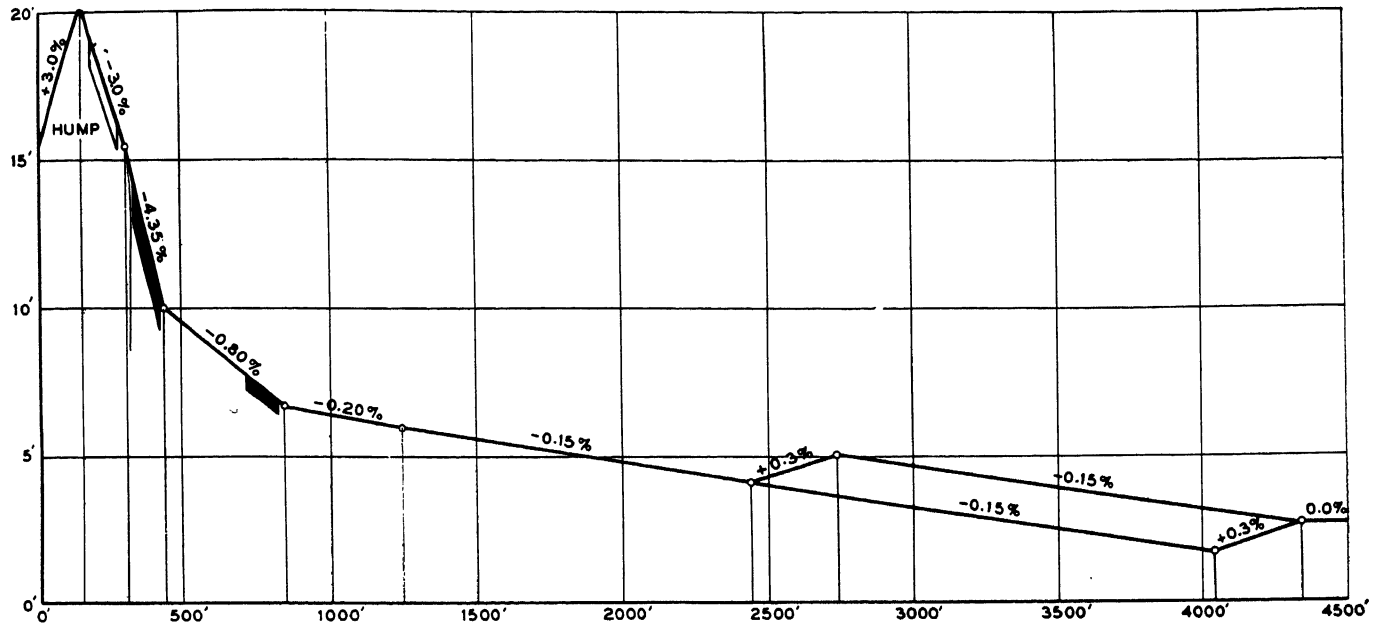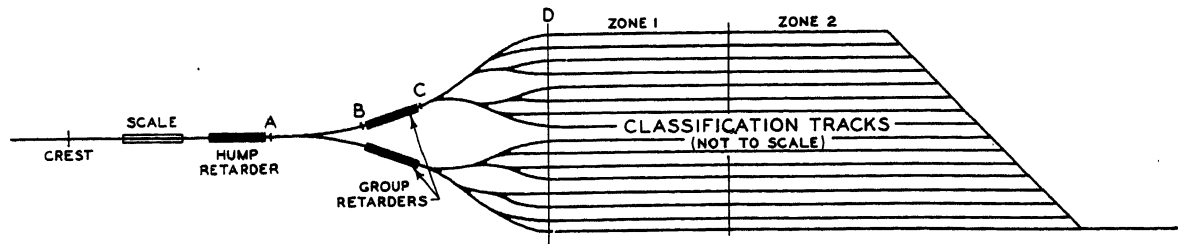
Fig. 2. Profile for gravity yard of type shown in Fig. 1

of conditions. He develops a pattern of release speeds which will avoid having cars overtake each other and go into the wrong track or couple at excessive speeds.

When the car weight information is reliable, manual control becomes an art in which the best operators develop a high degree of accuracy and consistency. To be economically justifiable, automatic operation must at least equal this performance.

A major step in the progress of yard automation was taken in December 1954, when the Elgin, Joliet & Eastern Railway Company at Kirk Yard, Gary, Ind., inaugurated a system in which the cars automatically establish and select their own releasing speeds derived from a computing system which has been fed the necessary elements of information.

Automatic Control Problems

As with manual control, there are two basic interrelated problems which must be solved in automatic control.

1. Provide sufficient separation between the cars to enable them to be switched to the tracks selected by the automatic switching system.

2. Predict rolling resistance of individual cars or cuts to establish releasing speeds from the last retarder which will ensure safe couplings on the classification tracks.

The following factors affect the speeds at which cars should be released from the retarders and therefore the solutions to the problems:

1. Factors associated with the yard design
   A. Amount of curvature in route which car will traverse
   B. Special track conditions in the route such as low spots, tight gauge, rough track, etc.
   C. Gradients along the route
2. Variable factors
   A. Special rail conditions such as rust, water, lubricants, etc.
   B. Wind direction and velocity
   C. Temperature
3. Factors associated with the car or cut of cars
   A. Over-all rolling resistance of car including that caused by center plates and side bearings
   B. Car lading
   C. Car weight
   D. Distance to coupling point

Solutions

The amount of separation required

between cuts is a function of the shortest length of track circuit which cannot be spanned by the trucks of the longest car. Separation is obtained by increasing the speed of cars moving down the grade above that at which the cars move toward the crest of the hump. Separation speed V_2 for the ideal case of uniform rolling resistance for all cars is
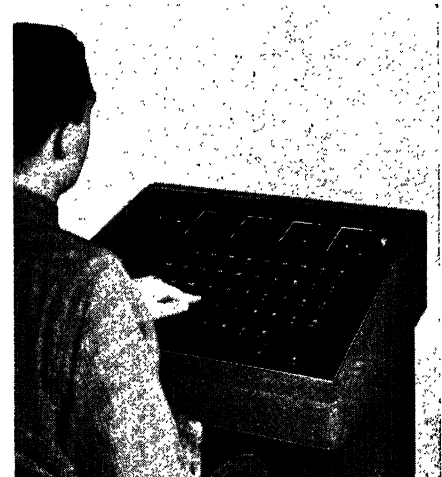
Fig. 3. Pushbutton panel for automatic selection of routes
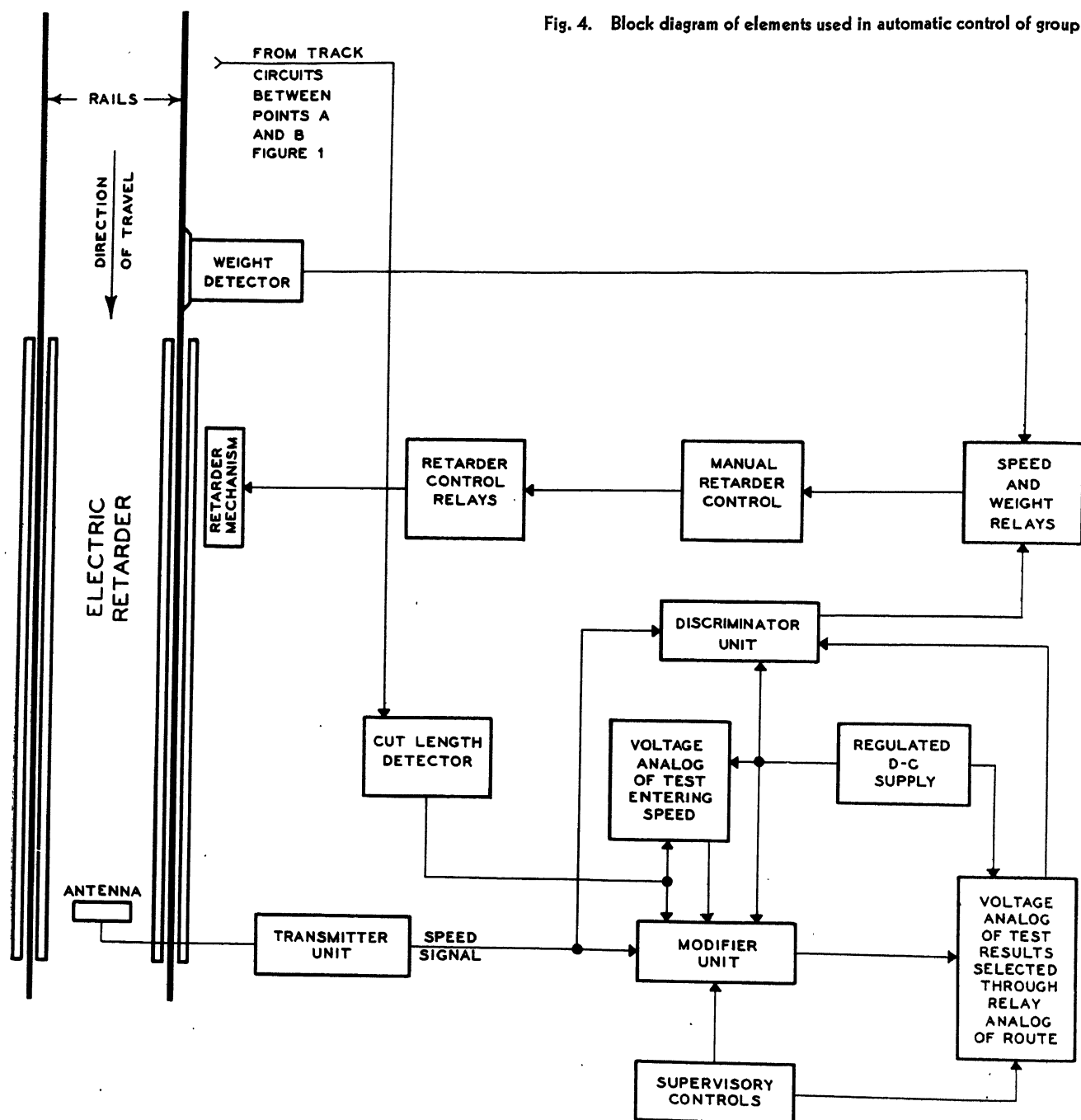
$$V_2 = V_1(S+L_t)/L_c$$

where V_1 is the velocity approaching the hump crest, L_t the length of the detector track circuit, L_c the length of the car, and S the distance between first and last axles of the car.

The humping speed V_1 is limited by V_2. When V_2 is the velocity leaving the last retarder, it in turn is limited by the maximum allowable coupling speed on the classification tracks. To keep V_2 and V_1 as high as possible, the grade beyond the last retarder must be non-accelerating. This establishes the maximum release speed for easy-running cars.

It is apparent that if there is a wide range in rolling resistance between cars there will be a tendency for easy running cars to overtake hard-running cars in the area between the hump crest and the group retarder. Conversely, hard-running cars, released at relatively higher speeds to compensate, will tend to overtake easy-running cars between the group retarder and the last switch in the route. These two factors tend to be mutually compensating. By using only two points of retardation and releasing all cars at a uniform rate from the first (hump) retarder, the detrimental effect on the maximum humping speed can be minimized.

In actual practice, the restrictions on humping rate which result from spacing requirements are usually alleviated by the random distribution of cars to the

Fig. 5. Electric retarder with radar antenna installed between rails
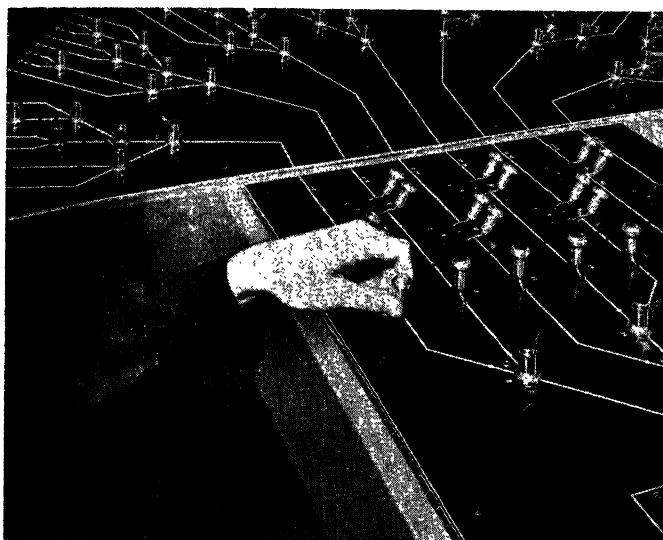
classification tracks. They do not in general follow adjacent routes, the most restrictive condition.

Uniform releasing speeds are obtained at the hump retarder by means of speed-measuring and fixed speed-selecting elements of the control system. Rolling resistance is measured and predicted in the following manner with the organization of elements shown for a group retarder in Fig. 4.

The characteristics of the tracks between points A-B and C-D in Fig. 1 are determined by means of a test car of low rolling resistance which is used to measure the average rolling resistance in those areas and on tangent track. This resistance includes curvature, switches, and any unusual local track conditions. Tests have shown that the ratio of the resistance ahead of the group retarder to that beyond, while usually different for each route, is relatively constant for cars having a wide range of rolling resistance. Therefore, if the average resistance between points A and B is measured for each cut of cars, the behavior beyond point C can be predicted within reasonable limits.

The releasing speeds for the test car required to produce safe coupling speeds at D are set up in terms of direct voltages on a resistance network analogue of the tracks beyond C. These serve as a reference for all other cars. Curves of releasing speeds are computed for cars which may range in rolling resistance from 0.2 to 1.4 per cent in equivalent grade.

When cars are released at constant velocity from the hump retarder, the average rolling resistance between the hump and group retarders may be expressed as

$$R = (-V_B^2/2g + K)/L$$

where V_B is the velocity of the cut as it reaches B, K a constant equal to the difference in elevation between A and B plus the elevation equivalent of V_A (the releasing speed at A), g is gravity, and L the length AB.

For the test car V_B can be made a reference speed and deviations from this reference will then be functions of R. Speed is established as a direct voltage, which is compared with a direct voltage proportional to the speed of each cut as it reaches point B. The resulting deviations are amplified (multiplied) by an adjustable amount in a modifier unit, and stored. The voltage increment is added to the voltage appearing on the analogue for the track so as to produce the voltage equivalent of the correct releasing speed. In cases where it is desired to make V_A variable for special purposes, similar results can be achieved by measuring V_A and transferring a correction to the reference speed at B.

Speed of cars is measured instantaneously and continuously as they approach and pass through retarders by means of 2,475-megacycle Doppler-type radar. The radar antennas are placed in the lower end of each retarder facing the oncoming cars; see Fig. 5. The Doppler frequency shift produces a difference frequency in the low audio range, which is directly proportional to car speed. It is fed from the transmitter unit to a discriminator where it is converted to a direct voltage proportional to velocity and compared with the voltage requested by the analogue network. If the voltage produced by the car is over that requested, the retarder is positioned to reduce the car speed. As this speed is approached, a partial release is made, followed by a full release when the selected speed is reached.

The rate of retardation is proportional to weight for a given retarder pressure. To secure uniform performance, the cars are placed in three weight classifications, light, medium, and heavy. These are obtained from an electromechanical device placed ahead of the retarder which weighs each wheel on one side of the cars and feeds the information to the control system. Retarder pressures are selected accordingly. These automatic speed and weight controls of retarder pressure are derived from relays in the output channels of the discriminator.

Rolling resistance measurements are affected by the length of the cuts. Because a long cut will roll free for a relatively shorter distance between positions A and B of speed measurements, than will a single car, it is evident that its entering speed will be less than for a single car of the same resistance since

Fig. 7. View of large modern yard equipped with automatic route selection and automatic control of car speeds

it is acted upon by the grade for a shorter interval. Compensation for this inherent source of error is obtained by classifying the cuts in three categories, short, medium, and long. This is accomplished by means of detector track circuits between points A and B. When a medium cut is detected, the reference entering speed at A is modified to correct for the fact that the test section was shorter. Medium cuts may be from two to four cars, and short cuts one to two cars. Cuts of five cars or more have such a short test interval that the differences in entering speeds approach the error factor in measurement. Therefore these select an arbitrary release speed which is adjustable and based on average values of rolling resistance.

Depth of penetration of the yard is controlled by the amplifying ratio in the modifier unit which computes the correction to the reference release speed. When conditions warrant, this can be an adjustment made automatically by counting cars entering each track. Usually a manual supervisory selection of two or three zones will provide sufficient adjustment for normal operations. This range adjustment applies only to cars which have a rolling resistance higher than the grade of the yard tracks. Cars having a

resistance equivalent to the grade are released at the same speed for all depths of penetration.

Conditions of wind, temperature, snow, and rain can modify the over-all operation. Compensation for changes brought about by these conditions is available on a control lever which provides adjustments to the normal release speed settings; see Fig. 6.

Manual control levers with four positions are provided. One position selects automatic operation and the others select retarder pressures which override automatic control. Associated with the group retarders are supervisory controls for selecting the zones of penetration.

Controls are also available for each power switch so that the operator can select routes for reverse moves from the classification yard to the hump.

Automatic Operation

Under normal conditions the string of cars being classified moves to the crest at constant speed and the retarder controls are in the automatic position. The retarder operator, located in a tower near the group retarders, watches the cars for unusual rolling conditions, and for errors

which may be made in the route selection. He does not touch the controls except to make occasional adjustments to the zone selector. In case of unusual conditions, such as a car with dragging brakes failure of cars to uncouple at the crest, or too fast a humping speed, he may use manual control until normal operations are restored. A large modern yard is shown in Fig. 7.

Summary

This system for automatically controlling the movement of cars into classification tracks is now operating in several classification yards and produces the following benefits:

1. More consistent and accurate control than can be obtained manually.

2. Reduction in damage to lading by minimizing unsafe couplings.

3. Maintenance of high humping speeds at all times.

4. Reduction in work strain on operator, which enables him to spend more time supervising the over-all operation.

5. Operations can be maintained during heavy fog conditions when cars are invisible to the operator.

Harmonics from Railroad Rectifiers on Power System Reduced by Filters

S. J. BOZZELLA
MEMBER AIEE

J. L. KENNEDY
NONMEMBER AIEE

M. MAHR, Jr.
NONMEMBER AIEE

H. W. WAHLQUIST
MEMBER AIEE

AT THE close of World War II, the Long Island Rail Road Company purchased the electrical equipment at the Aluminum Reduction Plant at Maspeth, N. Y., from the War Assets Administration. Included in this equipment were several 3,000-kw 6-phase 60-cycle mercury arc rectifiers, complete with the component assemblies which the Long Island Rail Road Company proposed to utilize in the expansion and modernization of their railroad electrification facilities on Long Island.

At the time that these rectifiers were purchased by the railroad company, the

Long Island Lighting Company had already begun construction of a new 13.2-kv distribution system. Since the rectifier transformers were also rated 13.2 kv on the primary side, it made it desirable from the power company's viewpoint to supply both the railroad load and the distribution load from a common 13.2-kv source. Difficulties associated with procuring substation sites and equipment, and the length of time for deliveries made it necessary to utilize available facilities to supply the railroad load.

During the early discussions between the railroad company and the power com-

pany, it was recognized that 6-phase rectifiers are a source of harmonic currents which tend to distort the power system wave shape. The New York Telephone Company and the Long Island Lighting Company operate under a joint-use contract whereby both utilities occupy poles jointly for their facilities. Considering the size of the rectifiers involved, use of a common 13.2-kv source to supply both railroad and distribution load would expose the telephone circuits to power-system harmonics having magnitudes sufficient to create serious noise problems in the telephone system.

The three utilities, recognizing the need for an early solution, appointed a committee with representatives from each utility to study the problem and to develop a filter which would reduce the harmonic voltages on the 13.2-kv supply circuits to a maximum sustained telephone influence factor (TIF) of 40. It was the opinion of the telephone company that with a TIF held to 40, the majority of subsequent noise problems on the telephone system could be taken care of by providing such remedial measures as

WEST BABYLON SUBSTA.
LONG ISLAND LIGHTING CO.
66 KV
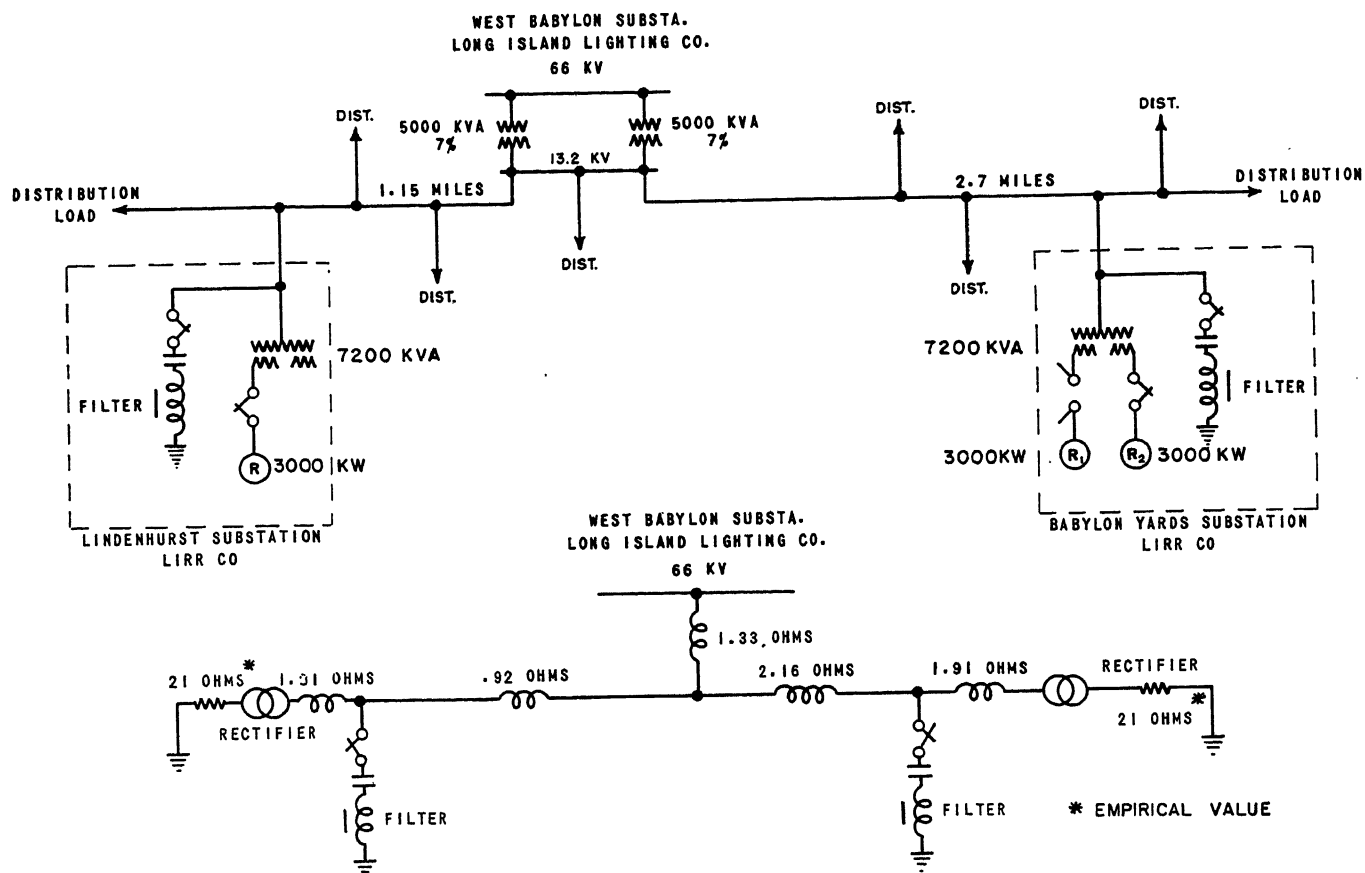
Fig. 1. Circuit impedances at 60 cycles

cable sheath shielding and improvement in system balance.

Because of the difficulty of securing material, particularly of special design, the committee was requested to develop a filter which would utilize standard equipment for its component parts. Filters were designed and installed at five locations. However, only the installations at the Babylon Yards rectifier substation and the Lindenhurst rectifier substation, shown diagrammatically in Fig. 1, are discussed in this paper, since the others are similar in design. The Babylon Yards substation is typical of a terminal station where steady car-heating loads are encountered in addition to load swings from normal train operation. Lindenhurst is

S. J. BOZZELLA is with the Long Island Lighting Company, Mineola, N. Y.; J. L. KENNEDY is with the Long Island Rail Road Company, Jamaica, N. Y.; M. MAHR, JR., is with the New York Telephone Company, Brooklyn, N. Y.; and H. W. WAHLQUIST is with Ebasco Services, Inc., New York, N. Y.

typical of a rectifier substation supplying only transient train loads.

Studies of the power system impedances involved in the proposed installations indicated that, to realize adequate reductions in the over-all TIF, the filters would have to be effective at all odd non-triple harmonics from the seventh (420 cycles) to the 35th (2,100 cycles). In addition, the studies showed that certain locations would require suppression of the fifth harmonic (300 cycles) to avoid substantial increases at this frequency due to resonance between the filter capacitance and system inductance. In view of the large magnitude of 300-cycle current generated by a 3,000-kw, 6-phase rectifier, the designing of suitable shunts for this frequency presented a number of difficulties.

Filter Design

Studies of various possible filter combinations indicated that the simplest and most economical arrangement would consist of three resonant shunts per phase, the individual shunts being tuned to 300, 420, and approximately 800 cycles.

In designing the filter, every effort was made to specify readily obtainable material for the component parts. An

investigation was made to determine the feasibility of modifying certain standard distribution transformers for use as reactors. The transformer selected was the Westinghouse Hypersil split-core 15-kva 1-phase 60-cycle type-S primary 480 volts with 2-5 per cent (%) taps, secondary 240/120 volts, style *1483646*, catalog no. *6UID-001*. The transformer tank and core and the winding assembly are shown in Fig. 2.

The core design of these transformers makes them especially adaptable for use as reactors since the split type of core enables the insertion of spacers to create an air gap, thereby providing a relatively constant inductance over a wide range of current. Micarta and varnished cambric having a thickness of 0.025 and 0.010 inch respectively were used for spacers.

From measurements made with various gap settings to obtain a series of magnetization curves, it was determined that a gap setting of approximately 0.117 inch provided satisfactory linearity and suitable inductance for the filter. The 60-cycle impedance values for a range of current are shown in Table I.

A variety of inductance values with the fixed gap setting are available by connecting the primary and secondary windings in series aiding, series opposing, parallel, or separately. Further varia-

WINDING	TAP	OHMS	APPROX. 60 CYCLE CURRENT RATING AMPERES
480 + 240	1	43	17
480 + 240	2	40	
480 + 240	3	37	
480 + 120	1	30	20
480 + 120	2	27.6	
480 + 120	3	25	
480	1	19	25
480	2	17	
480	3	15	
480 - 120	1	10.8	30
480 - 120	2	9.4	
480 - 120	3	8.1	
480 - 240	1	4.8	35
480 - 240	2	3.9	
480 - 240	3	3.1	
240	—	4.8	40
120	—	1.2	80

Fig. 2 (above). Modified 15-kva 480/240/120-volt transformer for use as iron core reactor
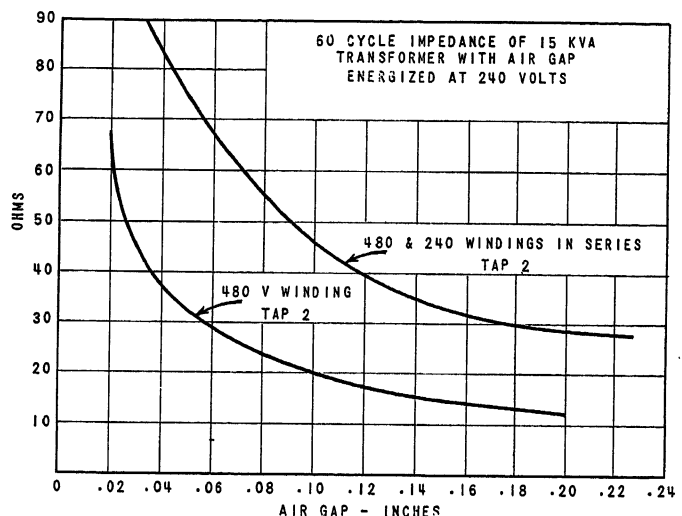
Fig. 3 (right). Sixty-cycle impedance combinations obtainable with 15-kva transformer with 0.117-inch air gap

tions are available by the use of two 5% taps on the 480-volt winding. The 60-cycle reactances and approximate current ratings for various winding combinations and a fixed gap of 0.117 inch are shown in Fig. 3. The wide range of values available in one unit· makes the reactors interchangeable, keeps the number of spares to a minimum, and facilitates tuning the shunt with fixed capacitors.

When these reactors were first placed

in service, hot spots developed in the transformer case opposite the air gap during heavy loads on the rectifier. These hot spots were created by eddy currents in the steel caused by concentration of flux opposite the air gap. To reduce the heating effects, a rectangular opening of 4 by 7 inches was cut in the case opposite the air gap and covered with a non-magnetic copper plate, as shown in Fig. 2.

The use of 15-kva transformers as reactors in the 300-cycle shunt at the Babylon Yards rectifier substation proved inadequate. Tests indicated that the filter functioned effectively for loads up to 3,000 kw. However, when train loads were superimposed on the 3,000-kw steady-heating load (heating in cars stored in the terminal station), swings in excess of 5,000 kw (9,000 amperes d-c) were observed. Under this condition, the 300-cycle iron-core reactors became saturated and no longer functioned with the capacitors to form a tuned shunt at 300 cycles. This caused resonance at 300 cycles between the filter capacitance and power system inductance, and the high current in the reactors caused them to fail because of excessive heating.

Because of the high loads encountered at the Babylon Yards rectifier terminal station, it appeared desirable to replace

the inadequate 15-kva iron-core reactors in the 300-cycle shunt with larger units or, to substitute air core reactors. To avoid long delay in delivery, air core reactors were fabricated in the lighting company shops utilizing standard no. 3 TBWP wire mounted on a Micarta frame. When this air core reactor was tested, it too was found to be inadequate since its design did not provide sufficient ventilation, and excessive heating resulted under heavy-load conditions. The air core reactors were modified with the use of vinyl-covered no. 2 American wire gauge wire with air spaces between every few layers and mounted on a transite core,

Table I. Sixty-Cycle Impedance of 15-Kva Transformer with 0.117-Inch Air Gap

Impressed Voltages	Current, Amperes	Impedance, Ohms
240-Volt Winding, D-C Resistance 0.025		
10	2.1	4.76
20	4.2	4.77
30	6.3	4.77
40	8.4	4.75
60	12.5	4.80
80	16.8	4.76
100	21.0	4.76
140	29.6	4.73
200	42.8	4.67
250	54.5	4.59
120-Volt Winding, D-C Resistance 0.0064		
10	8.4	1.19
80	68.0	1.18
480-Volt Winding, Tap 2, D-C Resistance 0.079		
40	2.3	17.4
200	11.5	17.4
300	17.3	17.3
500	30.4	16.4

Table II. Temperature Rise of 20-Ohm Air Core Reactor with 60-Cycle Current

60-Cycle Current,* Amperes	Duration of Test, Hours	Temperature Rise,† Degree Centigrade
18.5	1	1
31.3	2	8.5
41	2	23
51	1	34
60	1	58
60, Fan-Cooled	1	17.5

* In actual service these reactors are subjected to continuous current of 16.4 amperes at 60 cycles, with short-time current of 25 amperes at 300 cycles.
† Ambient temperature 23 to 28.
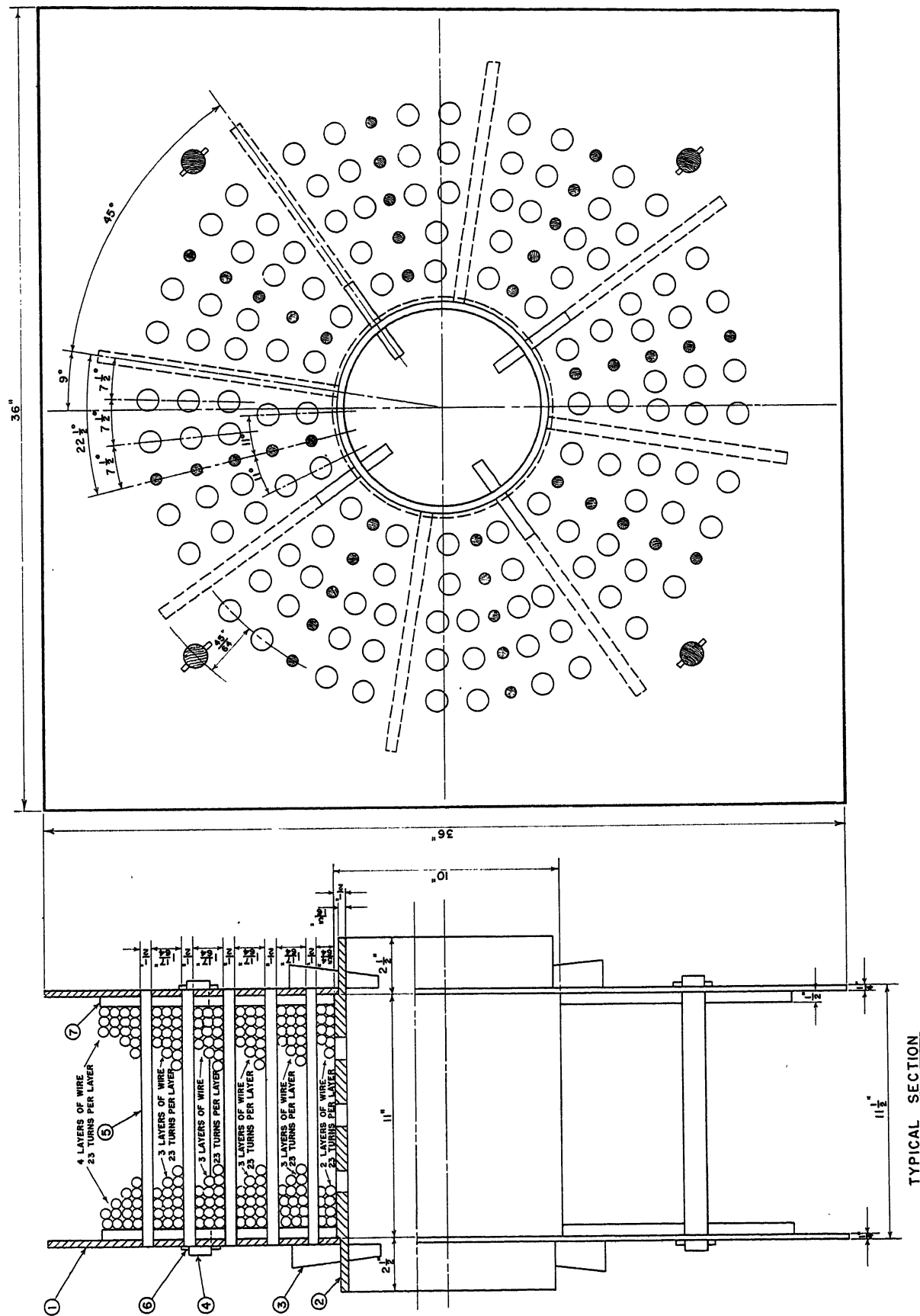
Fig. 4. Air core reactor design, 300-cycle shunt, Babylon Yards substation. Materials were as follows:

1—Micarta square 1/4 by 36 by 36 inches with 9⅝-inch-diameter hole
2—Transite cylinder 1/2- by 16- by 10-inch outside diameter
3—Wood wedge 1/2 by 4 inches tapered from 1 to 1/2 inch
4—1-inch wood dowel rods 13 inches long

5—1/2-inch wood dowel rods 12 inches long
6—1/4-inch wood dowel pegs 2 inches long
7—1/2- by 1/2- by 10³/₄-inch maple spokes
8—2,200-foot no. 2 Vinyl-insulated wire

TYPICAL SECTION

4 LAYERS OF WIRE 23 TURNS PER LAYER
3 LAYERS OF WIRE 23 TURNS PER LAYER
3 LAYERS OF WIRE 23 TURNS PER LAYER
3 LAYERS OF WIRE 23 TURNS PER LAYER
2 LAYERS OF WIRE 23 TURNS PER LAYER
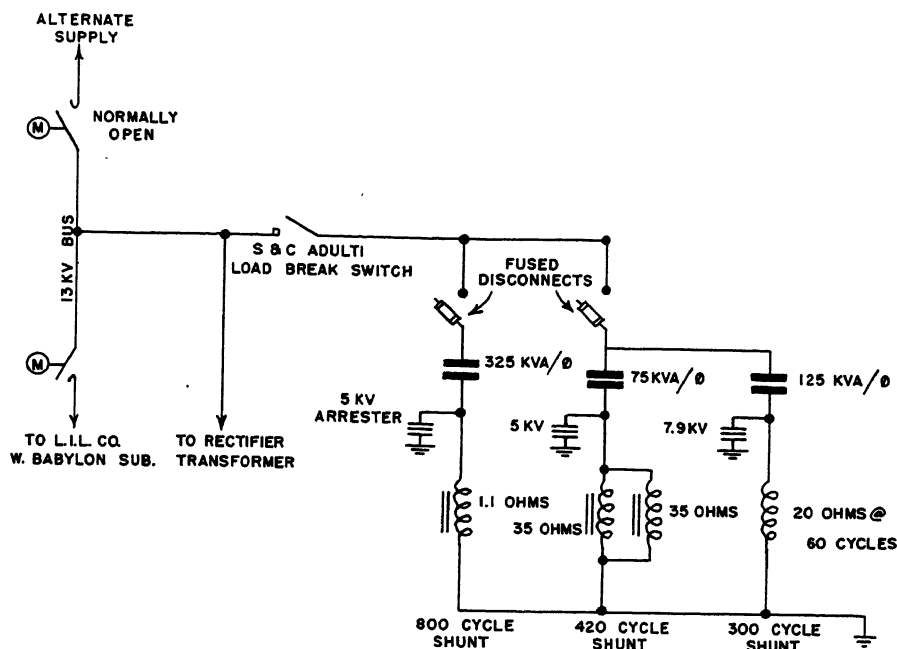
Fig. 5 (above). Harmonic filter, Long Island Rail Road, Babylon Yards substation
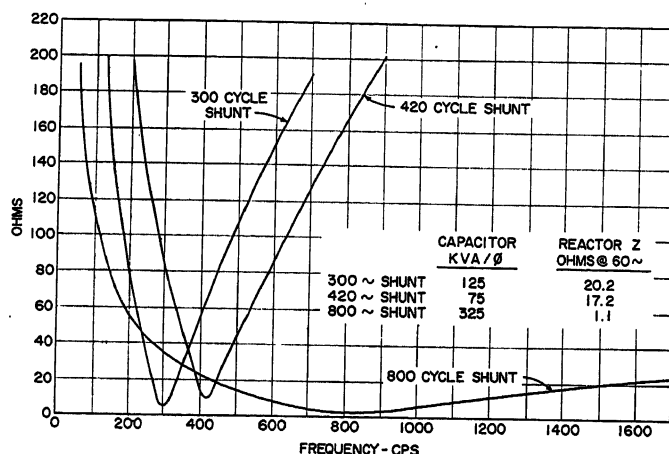
Fig. 6 (center left). Impedance of shunt elements, Babylon Yards filter
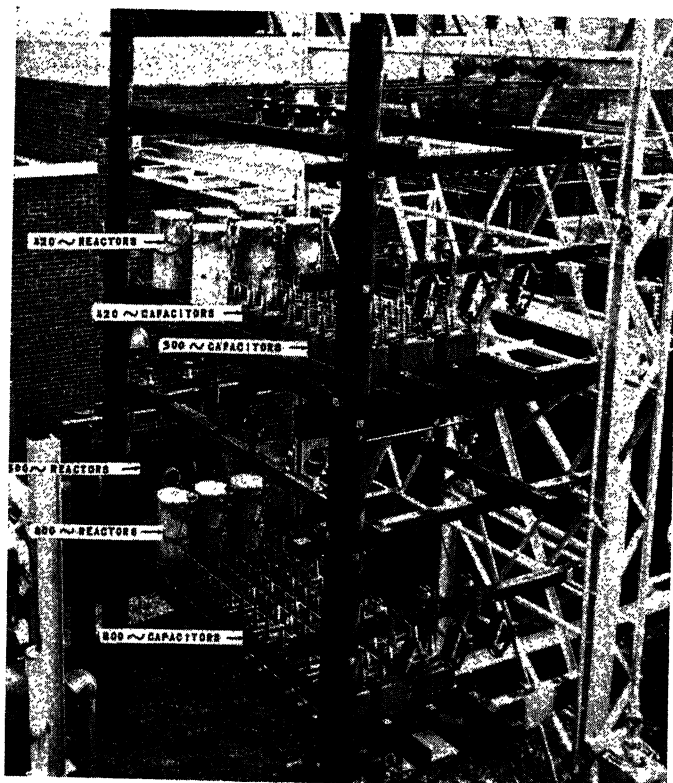
Fig. 7 (left). General view of harmonic filter installation at Babylon Yards Rail Road substation

as shown in Fig. 4. This latter design has proved satisfactory with no evidence of overheating. The temperature rise for a range in 60-cycle currents is shown in Table II.

At the Lindenhurst rectifier substation, 25-kva transformers with air gaps were used as reactors in the 300-cycle shunt. Although the shunts at Lindenhurst are not subjected to the steady heating loads, it appeared desirable to utilize the larger reactors in the 300-cycle shunt to provide a greater margin of safety.

As shown in Fig. 5, the bulk of the capacitance in the filter is used in the 800-cycle shunt (325 kva per phase) in order to realize sufficiently low shunting impedances to suppress all the harmonics in the range of 660 to 2,100 cycles. With this amount of kva, the harmonic loading of the capacitors in the 800-cycle shunt is negligible, and the low impedance of the associated reactor results in only 0.5% increase in the 60-cycle voltage across the capacitors. This, together with the non-critical nature of the tuning, provided conditions particularly favorable to the use of standard 25-kva power factor correction capacitors. The impedances of the shunt elements of the Babylon Yards filter over the frequency range of interest are shown in Fig. 6, and the installation is shown in Fig. 7.

In the design of the 300- and 420-cycle shunts, it was necessary to provide sufficient capacitive kva to handle safely the harmonic currents and voltages superimposed on the 60-cycle values. The industry standards for power factor correction capacitors allow continuous kilovar (kvar) loading of 135% for combined 60 cycles and harmonics. Harmonic analyses with experimental shunts in service had shown the maximum magnitudes of 300- and 420-cycle shunt currents expected during peak loads were in the order of 25 amperes at 300 cycles and 10 amperes at 420 cycles. Using 75 kva of capacitors per phase in the 420-cycle shunt, the 420-cycle kvar at 10 amperes amounts to only 16% of the 60-cycle kvar. The choice of capacitor kva rating for this frequency was therefore governed by considerations of the shunt impedance necessary to realize adequate effectiveness at the tuned frequency rather than harmonic loading.

With 125 kva of capacitors per phase in the 300-cycle shunt, the harmonic kvar with 25 amperes of 300 cycles amounts to 50%. This, combined with a 4% increase in 60-cycle voltage, results in a total kvar of 198 or 158% which is substantially above the industry rating for continuous duty. However, since these maximum

Bozzella, Kennedy, Mahr, Wahlquist—Harmonics Reduced by Filters   NOVEMBER 1955

values of harmonic current occur for short periods only during combined car-heating and train-starting loads, no appreciable heating of the capacitors results. In view of this, and since 125 kva per phase is adequate from a harmonic-suppression standpoint, the use of higher capacitor ratings is not justified. All of the shunts therefore employ standard 25-kva power factor correction capacitor units which not only simplifies replacement and reduces outage time, but also avoids the necessity of stocking capacitors with special voltage ratings. Detailed constants of the filter are shown in Table III.

A consideration in the use of standard power factor correction capacitors in filter designs is the variation in capacitance with temperature. Sharply tuned filters require oil-filled capacitors rather than chlorinated diphenyl (askarel)-filled units commonly employed for power factor correction. This is particularly important when the temperature of the filter capacitors may drop below 30 degrees Fahrenheit. For the filter under consideration, the broad tuning of the 800-cycle shunt avoids any appreciable influence on the filter effectiveness due to temperature changes. For the 300- and 420-cycle shunts, however, which have much higher inductance-capacitance ratios, appreciable change occurs in the tuned frequencies with capacitor temperatures below freezing. So long as the filter is continuously energized, the temperature changes are not sufficient to affect filter effectiveness seriously. The effects are partly compensated for by the increased losses that occur in the dielectric as the temperature drops. It is desirable to have the tuning correct for temperatures above 50 degrees Fahrenheit so that any reduction occurring in the capacitance at low temperatures shifts the tuning point of each shunt to a somewhat higher frequency. This ensures that the impedance of the three shunts in parallel will not undergo much change at 300 or 420 cycles, since it avoids the possibility of parallel resonance between shunts. Such resonance might occur if either the 300- or 420-cycle shunts becomes inductive at the nominal tuned frequency. Power factor correction capacitors as currently manufactured employ modified askarels which greatly reduce the variations in capacitance with temperatures as compared with early units.

In general, the constants of the filter with 300-, 420- and 800-cycle shunts are such that adverse resonant conditions in the balanced circuit due to the filter are very unlikely, regardless of power system impedance conditions.

In the residual circuit, it is possible for a resonant condition to result at 180 cycles under certain power system impedance conditions where the shunts are connected phase to ground. Should such resonance give rise to objectionable values of 180-cycle residual current, an additional reactor in series with the filter neutral to ground can be used to destroy the resonant condition. The reactor should have a 60-cycle impedance in the order of 10% of the 60-cycle impedance-to-residuals of the 300-, 420-, and 800-cycle shunts in parallel. For the filter such as that used at Babylon Yards, a reactor of about 4 ohms at 60 cycles would be required. From Fig. 3, the 240-volt winding of a 15-kva transformer with 0.117 air gap could be used.

Table III. Filter Constants, Babylon Yards

	300-Cycle Shunt	420-Cycle Shunt	800-Cycle Shunt
Total capacitor, kva	375	225	975
Capacitor kva per phase	125	75	325
25 kva capacitors per phase	5	3	13
Voltage rating of capacitors, volts	7,960	7,960	7,960
Shunt current at 60 cycles, amperes	16.4	9.65	41
Phase-to-neutral capacitance, microfarads	5.23	3.14	13.6
Reactor impedance at 60 cycles, ohms	20.2	17.2	1.1
Type of reactor	air core	iron core	iron core
D-c resistance of reactor, ohms	0.35	0.10	0.0064
A-c resistance at tuned frequency, approximate ohms	2.5	9	1
Reactor Q at tuned frequency	40	13	14
Reactor weight, pounds per phase	625	300*	300
60-cycle voltage across reactor, volts	331	166	45
60-cycle voltage across capacitor, volts	8,300	8,120	8,000
Effective resistance of shunt at tuned frequency, ohms	3	10	1
Reactor impedance at tuned frequency, ohms	101	120	14.6
Capacitor impedance at 60 cycles, ohms	506	843	195
Shunt impedance at 60 cycles, ohms	486	826	194
Effective 3-phase capacitive kva of shunt	392	230	980

* Weight of each, two used per phase.

Table IV. Voltage and Current Analyses on 13.2-Kv Supply to 3,000-Kw Rectifier at Babylon Yards

Filter on with Steady 4,500-Ampere Train-Heating Load and with Train Operation Swings up to 9,000 Amperes D-C

Frequency	Phase-to-Ground,* Volts	Voltage, TIF	Filter Current, Amperes*			Impedance of Shunts*		
			300-Cycle Shunt	420-Cycle Shunt	800-Cycle Shunt	300-Cycle Shunt	420-Cycle Shunt	800-Cycle Shunt
60	7,830	1	16.3	9.6	42	−478	−815	−188
180	43	0.3	0.5	0.24	0.97	−96	−220	−62
300	108	6.7	24	1.5	4.3	4.3	−74	−30
420	76	11.7	1.1	7.7	4.7	69	9.3	−18
660	47	18.9	0.30	0.55	9.7	160	87	−4.9
780	8.1	4.4	0.039	0.031	4.3	215	170	1.8
1,020	11.4	9	0.039	0.051	1.6	300	225	7.2
1,140, not measured								
1,380	15.5	13.2	0.038	0.058	0.87	400	270	17
1,500, not measured								
1,740	16	15	0.031	0.035	0.63	500	450	25
1,860, not measured								
2,100	10.9	10.7	0.017	0.02	0.28	640	540	32
RSS	7,830	34 40†	29	12.5	44			

* Average for phases A, B, and C.
† RSS value assuming TIF contributions of 12 for frequencies not measured.

Table V. TIF Contribution of Harmonics in 13.2-Kv Supply to Railroad Rectifier Substation, Babylon, N. Y., June 12, 1952

Harmonic	Frequency	TIF Contribution	
		Filter On*	Filter Off
5	300	2	39.8
7	420	4	14.1
11	660	7	26.6
13	780	13	35.5
17	1,020	8.5	39.8
19	1,140	13.3	31.6
23	1,380	6.3	68.4
25	1,500	4	37.6
29	1,740	10	47.3
31	1,860	5.6	33.5
35	2,100	7.5	37.6
37	2,220	5.0	22.4
RSS		25	100

* 300-, 420- and 800-cycle shunts connected.

Table VI. TIF Contribution of Harmonics in 13.2-Kv Supply to Railroad Rectifier Substation, Lindenhurst, N. Y.

Harmonic	Frequency	Shunts in Operation			
		300, 420, 800 TIF	420, 800 TIF	800 TIF	No TIF
5	300	2.5	21	25	11
7	420	1.5	5	45	13
11	660	8	16	13	34
13	780	7	8	10	50
17	1,020	6	10	9	115
19	1,140	7	7	14	35
23	1,380	8	14	18	40
25	1,500	10	11	8	36
29	1,740	8	18	5	47
31	1,860	7	8	4	36
35	2,100	5			45
37	2,220	7			
RSS		20	40	63	120
Measured total TIF		21	38	63	

Harmonic Analyses and TIF

Table IV shows the results of harmonic analyses on the 13.2-kv supply to the 3,000-kw rectifier at the Babylon Yards substation with filter in service. The measurements were made with a steady heating load of 4,500 amperes d-c at 600 volts combined with train operation loads which resulted in maximum short-time d-c currents of 9,000 amperes (5,400 kw). The controlling harmonic frequency in the voltage wave shape from the standpoint of magnitude is 300 cycles, amounting to 108 volts, or 1.4% of the 60-cycle fundamental. The controlling frequency in

the TIF is the 660-cycle component (11th harmonic) which contributes 18.9 to the total TIF. The controlling current in the 300-cycle shunt is 24 amperes at 300 cycles, the magnitude of other harmonics being small. Similarly, in the 420-cycle shunt, the tuned frequency component controls, the other harmonics being small. In the 800-cycle shunt, the 5th, 7th, 11th, and 13th harmonic components are prominent. The shunt impedances shown in Table IV were derived from the ratios of phase-to-ground voltages and shunt currents of the individual harmonic components.

Table V shows the TIF contributions of harmonics in the phase-to-ground voltage on the 13.2-kv supply to the Babylon Yards rectifier. These measurements were made with normal train operation loads under "filter on" and "filter off" conditions. The over-all root-sum-square (RSS) total of the individual components shows TIF values of 100 with the filter disconnected and 25 with the filter connected, which is a reduction of four to one.

Similar TIF contribution data are shown in Table VI for several filter combinations at the Lindenhurst railroad rectifier substation. With the 300-, 420-, and 800-cycle shunts in operation, the calculated total TIF with normal train operation loads is 20. Disconnecting the 300-cycle shunt increases the 300-cycle component by a factor of eight, and the total TIF by a factor of two. With only the 800-cycle shunt connected, large increases occur at 300 and 420 cycles, resulting in a total TIF of 63. With the filter off, the TIF increases to 120.

Under certain types of system impedance conditions, the increases in 300- and 420-cycle components are much larger than shown in this table when the 800-cycle shunt alone is used. At Babylon Yards, for example, disconnecting the 300- and 420-cycle shunts results in parallel resonance between the 800-cycle shunt
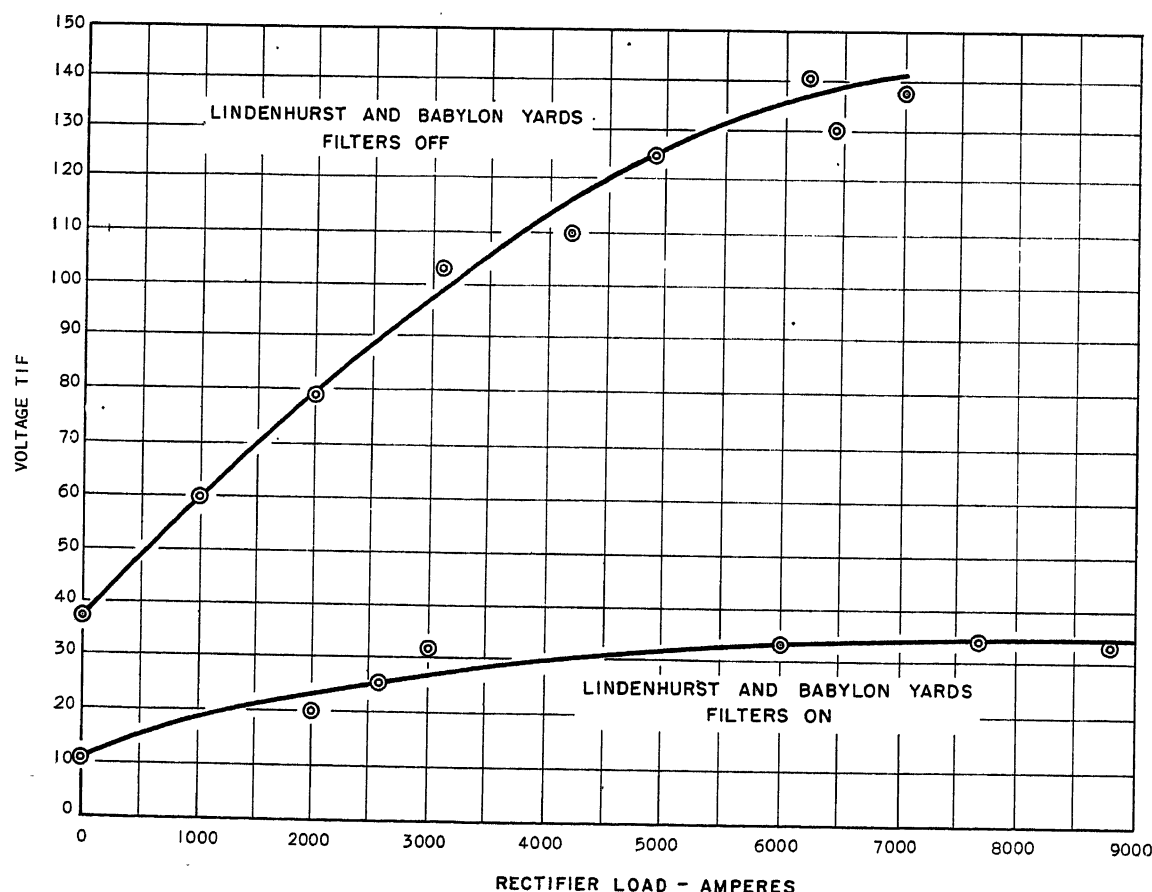
Fig. 8. TIF versus rectifier load current at Babylon Yards, February 2, 1954

Bozzella, Kennedy, Mahr, Wahlquist—Harmonics Reduced by Filters

TIF VS RECTIFIER LOAD CURRENT
AT LONG ISLAND RAILROAD SUBSTATION
PORT WASHINGTON, NEW YORK
JULY 18, 1952
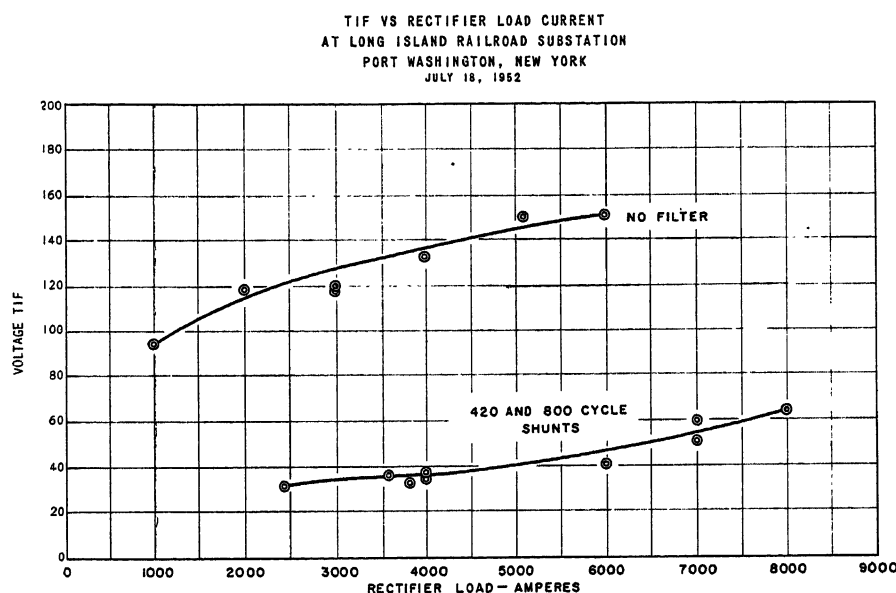
capacitance and the system inductance close to 300 cycles. Under these conditions, the TIF contributions at 300 and 420 cycles alone are approximately 100. At some other locations, the system impedances are sufficiently low to permit operation with only the 800- and 420-cycle shunts, without causing much increase in the 300-cycle component.

The curves in Figs. 8 through 10 show the variation in voltage TIF as a function of d-c load on the rectifiers at three of the stations where filters are installed. The Babylon Yards and Lindenhurst filters employ 300-, 420-, and 800-cycle shunts, whereas the Great Neck and Port Washington filters employ only 420- and 800-cycle shunts.

Recording Meter Measurements

The extreme variations in rectifier load and TIF encountered in the type of electrified railway system involved in this investigation are illustrated by the graphs in Figs. 11 through 13, obtained with recording meters. The peak values occur when a train is accelerating near the rectifier station. In addition to the general pattern of peaks over a 24-hour period, there are day-to-day variations in TIF magnitudes, particularly for the "filter off" condition due to changes in system impedance with switching operations on the power system.
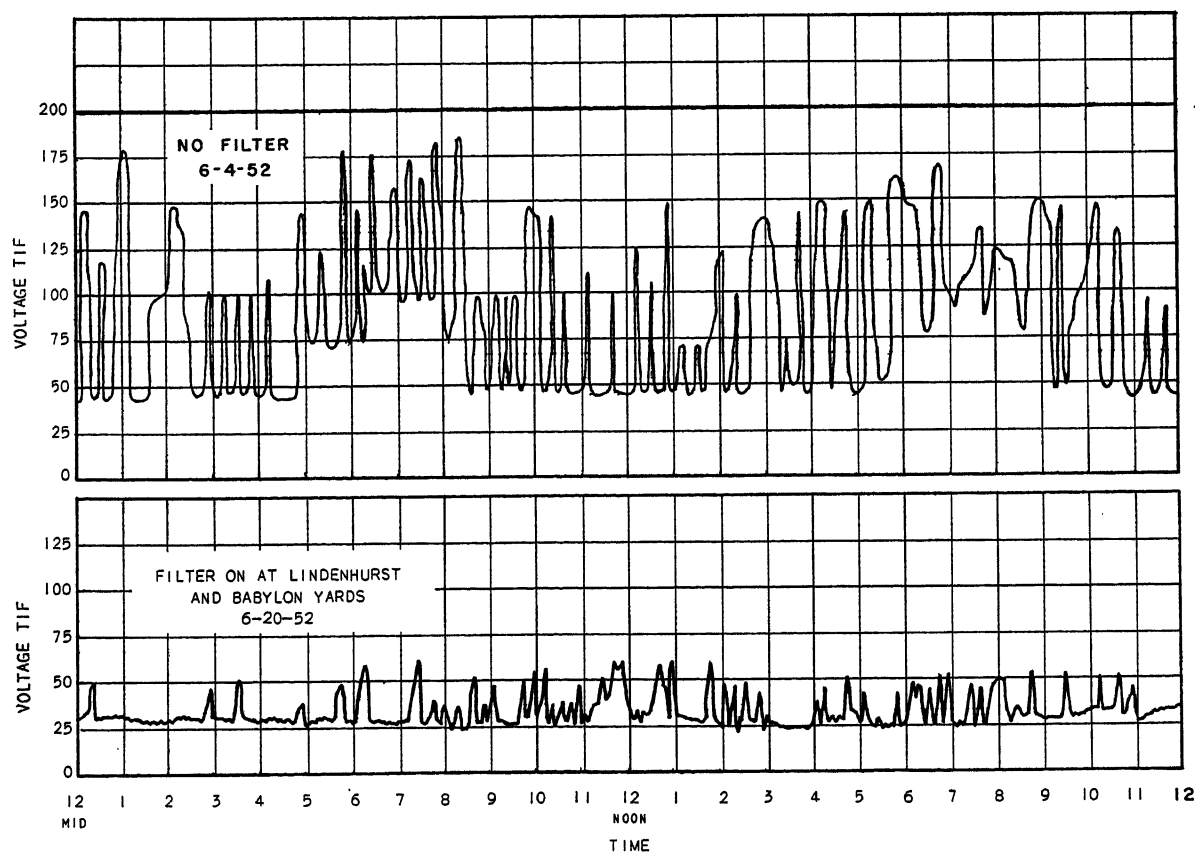
Fig. 12. Railroad rectifier substation. Babylon Yards charts represent envelope of maximum points
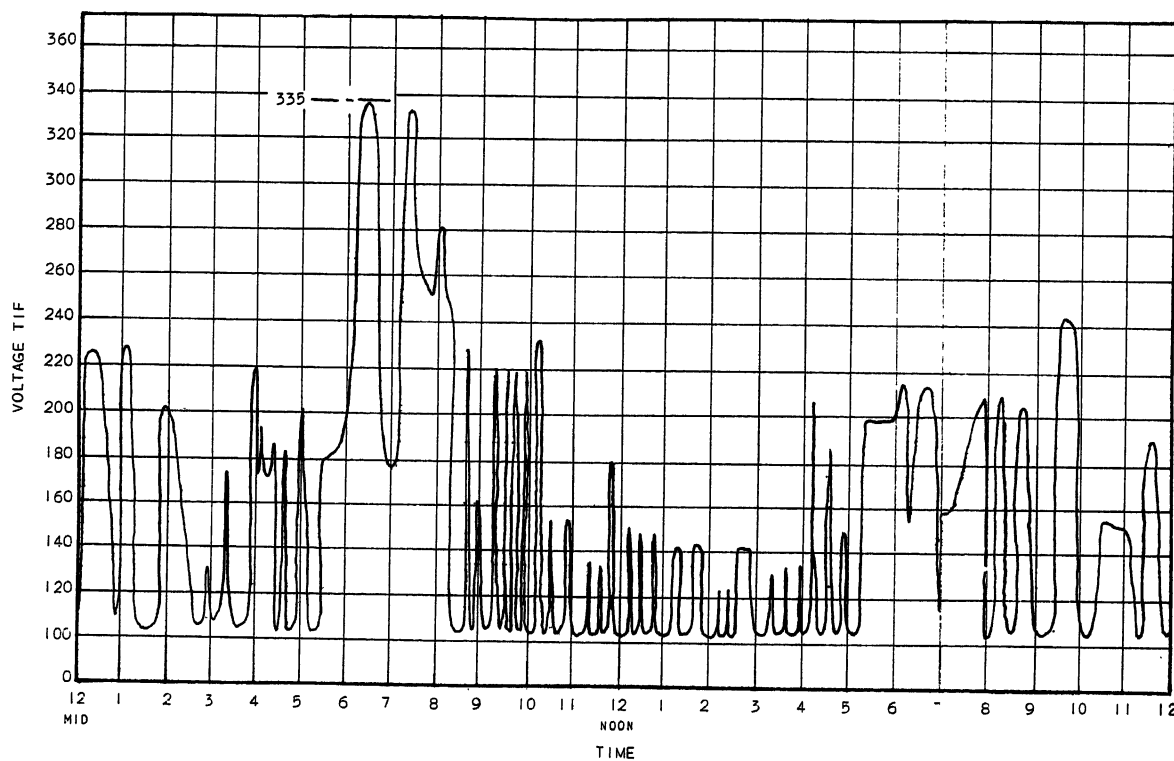
Fig. 13. TIF record. No-filter railroad rectifier substation, Babylon Yards. Envelope of maximum points of recording during 24-hour period

Expressions for Filter Effectiveness

The effect of a shunt filter on the power system voltage at a particular frequency is determined by the impedance of the filter, and the impedance of the power system looking in the two directions from the point at which the filter is located. The equations in connection with Fig. 14 are effectiveness factors considering all voltage sources on the system and they have a common numerator equal to the impedance of the shunt in series with the driving point impedance of the power system network. Where the impedances at a particular frequency are such that the voltage or current ratios are less than unity increases occur, and for ratios exceeding unity decreases occur, when the shunt is connected.

EFFECT OF FILTER ON LINE VOLTAGE, FIG. 14

Equation 1 applies for voltage at a given frequency that may originate at any point on the system; see Fig. 14.

Filter effectiveness $= \dfrac{V}{V_F} = \dfrac{Z_F + Z_{AB}}{Z_F}$ (1)

where

$V =$ voltage at point X, filter off

Bozzella, Kennedy, Mahr, Wahlquist—Harmonics Reduced by Filters
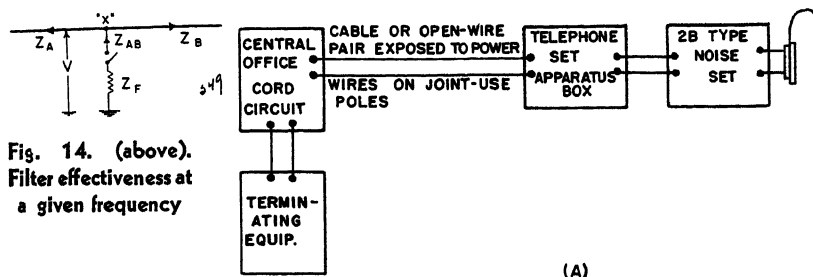
Fig. 14. (above). Filter effectiveness at a given frequency
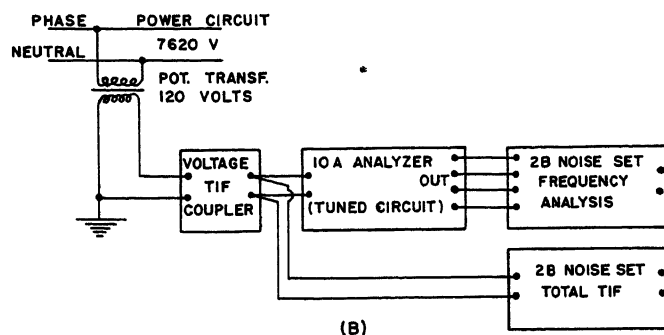
Fig. 15 (right). Equipment used for noise and TIF measurements

A—For noise measurements
B—For TIF measurements

(A)

(B)

V_F =voltage at point X, filter on
Z_F =impedance of filter
Z_A =system impedance in direction A
Z_B =system impedance in direction B
$Z_{AB}=Z_A$ and Z_B in parallel $=(Z_A \times Z_B)/(Z_A+Z_B)$

EFFECT OF FILTER ON LINE CURRENTS

In the current ratios of equations 2 through 4 it is necessary to take into account the source of the harmonic with respect to the shunt location.

For harmonic sources on A side of X

$$\frac{I_A}{I_A'}=\frac{Z_F+Z_{AB}}{Z_F+Z_B} \qquad (2)$$

$$\frac{I_B}{I_B'}=\frac{Z_F+Z_{AB}}{Z_F} \qquad (3)$$

For harmonic sources on B side of X

$$\frac{I_A}{I_A'}=\frac{Z_F+Z_{AB}}{Z_F} \qquad (4)$$

$$\frac{I_B}{I_B'}=\frac{Z_F+Z_{AB}}{Z_F+Z_A} \qquad (5)$$

where
I_A and I_B =line currents on sides A and B, filter off
I_A' and I_B' =line currents on sides A and B, filter on

Strictly, these relationships express the changes in voltage and current only at a given pair of terminals on the network. A shunt from the phase wire to neutral of a single-phase multigrounded line would represent a simple example of conditions where they have direct application. Practically, however, they are useful in analyzing the effects on the more complicated composite 3-phase and single-phase systems although special interpretation is required in dealing with certain of the residual and ground return components. The relations apply directly

to the effects of a wye- or delta-connected shunt on the balanced components of a uniform line when associated with the impedance of the balanced circuit. They also give the effect of a wye-grounded shunt on the residual (3×0 sequence) components which are directly impressed on the line when associated with the impedances of the residual circuit.

Measurements on Telephone System

Telephone outside plant in the areas involved in these tests consists of aerial cable and some open wire. The majority of the exposures involve 13.2-kv open-wire power distribution lines and telephone cables on joint-use poles. Exposures range up to 3 miles in length. Telephone central offices at Babylon and Lindenhurst are manual switchboard offices.

Standard Bell System practices were followed covering procedures in making noise measurements on local exchange cable telephone circuits. A Western Electric 2B noise set was used to measure telephone set receiver noise (HA-1 receiver) with connections as shown in Fig. 15(A).

Noise measurements were generally made during heavy railroad rectifier load hours of the early morning with the use of a test set that permitted tests on various individual line and party line antisidetone telephone station sets. The 2B noise set was connected to receiver terminals brought out to binding posts on the telephone set apparatus box, which provides switching to simulate the various types of station sets in use.

Receiver noise measurements were made at 16 locations in the central office areas of Babylon and Lindenhurst. Re-
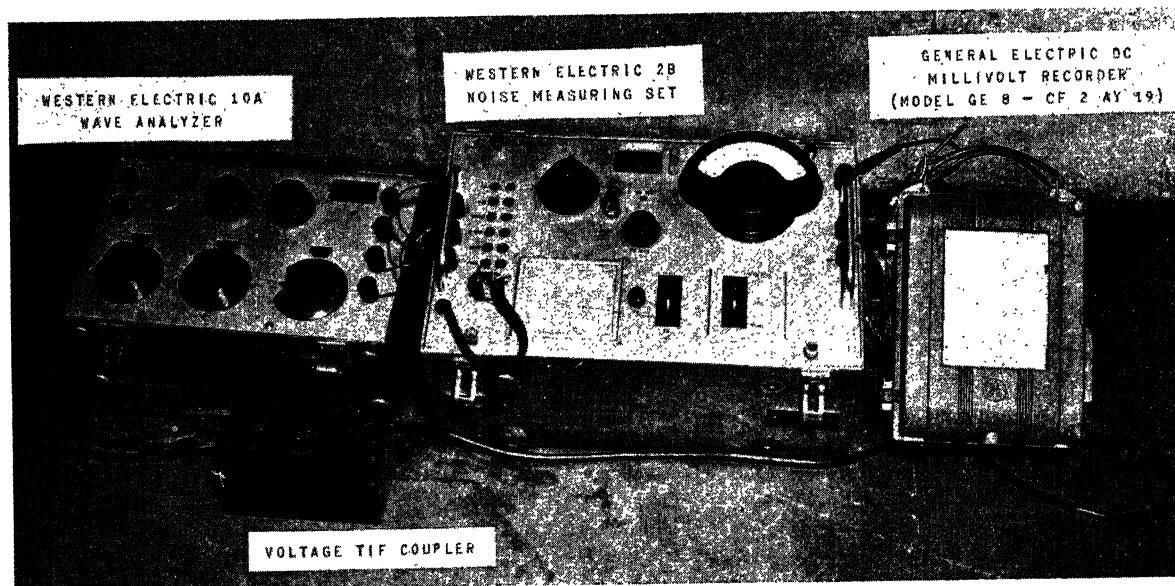
Fig. 16. Equipment used for wave shape analysis and TIF recording

SINGLE FREQUENCY TIF VALUES

FREQUENCY	TIF	FREQUENCY	TIF	FREQUENCY	TIF	FREQUENCY	TIF
60	0.5	1020	6090	1860	7380	3000	8420
180	61	1080	6190	1980	7580	3180	8340
300	479	1140	6310	2100	7770	3300	8060
360	804	1260	6510	2160	7850	3540	6880
420	1190	1380	6730	2220	7940	3660	6040
540	2090	1440	6820	2340	8070	3900	4520
660	3110	1500	6900	2460	8170	4020	3640
720	3630	1620	7120	2580	8220	4260	2280
780	4170	1740	7220	2820	8380	4380	1760
900	5160	1800	7300	2940	8450	5000	423
1000	6000						
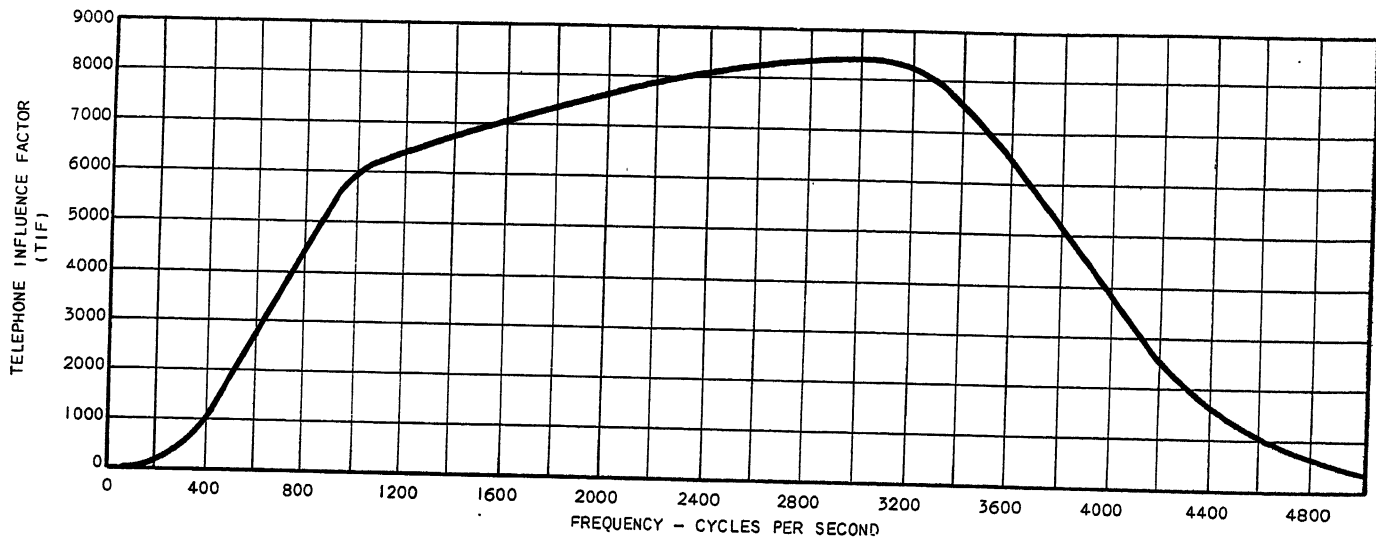
Fig. 17. Tentative TIF weighting based on 1941 FIA line weighting and 1,000-cycle TIF of 6,000

ceiver noise with the filters in use, measured at various terminals in cables with long exposures, are: Babylon, 8 to 17 adjusted decibels (dba); Lindenhurst, 8 to 18 dba. These levels have negligible effect on telephone circuits, a maximum of 0.5 decibels (db) noise transmission impairment (NTI). Appropriate corrections are used in the reading in dba, to compare the *HA-1* receiver response and accoustic efficiency to the *144* receiver. The value of 90 decibels below 1 milliwatt of power at 1,000 cycles in 600 ohms is the reference for *144* receiver noise measurements. The present, more commonly used, *HA-1* receiver has 5 db lower response at 1,000 cycles, and therefore 85 decibels below 1 milliwatt reference under the same conditions.

Telephone circuit noise can, if desired, be expressed in approximate equivalent db loss of transmission (NTI). The equivalent circuit loss due to noise can

be expressed in db to obtain the effect on a circuit that can be added to other loss values to determine an over-all circuit loss. The effect of the noise cannot be eliminated by providing a db gain equal to the NTI since it would also raise the noise level with the desired signal. The use of NTI is not acceptable to all engineers, though it may prove useful in certain cases.

Measurements of TIF

In Fig. 15(B), illustrating schematically the equipment used in measuring TIF

$$TIF = \frac{\sqrt{(V_a W_a)^2 + (V_b W_b)^2 + \ldots (V_n W_n)^2}}{V_{total}}$$

where

V_{total} = effective values of entire wave
$V_a, V_b, \ldots V_n$ = voltage of individual harmonics
$W_a, W_b, \ldots W_n$ = weighting of individual harmonics

Two *2B* noise sets were used to obtain the total TIF and that of individual components, making up the total TIF. Analyses of voltage were made on the 120-volt secondary of station-type potential transformers. Analyses of current were obtained from the voltage drop across a 0.5-millihenry shunt connected in series with the secondary of station-type current transformers. The TIF-measuring equipment was also used to operate a recording meter when TIF records of 24 hours or longer were desired. The TIF data on Figs. 11 through 13 were obtained in this manner. Fig. 16 shows the equipment used for analysis and TIF recording. All TIF values in this report are based on the 1941 TIF curve shown in Fig. 17.

Reference

1. ENGINEERING REPORTS OF THE JOINT D AND R SUBCOMMITTEE, Edison Electric Institute-Bell Telephone System, New York, N. Y., vol. I-V.

Discussion

C. W. Frick (General Electric Company, Schenectady, N. Y.): The paper gives an interesting example of the solution of one kind of inductive co-ordination problem. This problem may be considered as made up of two parts. The first part was to determine the type of solution and the second part was to reduce that type of solution to practical form.

There is a statement at the beginning of the paper that the three utilities appointed a committee to study the problem and develop a filter. The filter is not the only type of solution and the early decision to use filtering probably resulted from previous investigations over a period of years under the guidance of the Joint Subcommittee on Development and Research of the Edison Electric Institute and the Bell Telephone System. As an illustration of different types of solutions, the rectifiers had previously been used in an aluminum plant, and the 6-phase units were connected through phase-shifting transformers to ob-

tain multiphase operation. Certain harmonics can be suppressed by using multiphase arrangements such as 36 phases, as discussed in one of the Engineering Reports of the Joint Subcommittee. This scheme has advantages over filters but could not be used when the component rectifiers were installed at different locations after the railroad company had acquired the equipment.

Having decided to use filters, it remained for the committee to work out the details. There is a background of experience with filters for rectifiers of comparable kilowatt rating which dates back to about 1935. Factory-built units usually have the capacitors, reactors, and protective devices assembled in a tank similar to a transformer tank with three terminals which are conveniently connected to the 3-phase supply line at the rectifier input. Reactors are mostly of the air-core type. Separate tuned circuits are usually provided for harmonics up to 1,140 cycles. The filter as developed by the committee and described is not encased. It uses mostly iron-core reactors and includes one tuned

circuit to cover several harmonics. It has the advantage of construction on a short-time schedule, but it is not compact and connections are exposed. The details may be useful in future situations where short-time schedules are essential. However, factory-built filters are to be preferred when time permits.

There is need for both the long-range type of organization such as the Joint Subcommittee on Development and Research which has been in existence since about 1921, and organizations of the task group type to handle special problems such as this. Work of the continuing organization helps the task groups to become oriented and saves time. Furthermore, the long-range or continuing organization is needed to keep up with progress. For example, it is reported that the Joint Subcommittee on Development and Research is taking steps to bring the TIF weighting curve up to date and thus remove the necessity for using special data such as the tentative 1941 TIF curve used by the authors of the paper. It is hoped that the Joint Subcommittee will be given adequate support in this.

Problems Associated with the Development of a Power System for a Manufacturing Plant

W. C. HEINZ
ASSOCIATE MEMBER AIEE

GENERAL ground rules relative to the design of a power system for a manufacturing operation have been accepted and applied throughout industry. Each application must, of course, be examined with respect to its individual requirements and considerations. This is exemplified in the development of the power system for the new plant of the General Electric Company at Bloomington, Ill.

The facilities of the plant are utilized for the manufacturing, testing, and development of general purpose control

items such as magnetic motor starters, compensators, push buttons, and contactors. The plant is of modern single-story, rectangular contruction, with a manufacturing and testing area of approximately 326,000 square feet. An attached office section adds approximately 58,000 square feet to the working area.

This paper concerns the design and development of the power system for the Bloomington plant, and emphasizes some of the unusual problems encountered and their solution. The problems evolve from the requirement that the power system be designed to satisfy the general requirements of a typical manufacturing plant in addition to furnishing power for highly specialized testing in the development laboratory. The two loads have very different characteristics in that the testing requirements result in current demand of large magnitude lasting only a fraction of a second per test, while the general requirements in the plant are more constant.

Main Substation and Primary Distribution

The new plant in itself represents quite a sizable block of power for the utility circuits feeding the area. A study of the expected power requirements for the plant concluded that 5,000 kva for the main substation transformer would be sufficient to satisfy the needs of the plant for the present and the foreseeable future.

To reduce the adverse effects of the peaking-type welding and product development test loads on the normal plant system, several main substation arrangements were considered, including the use of: 1. a larger kva rating for the main substation transformer than would normally be required to supply the plant load, 2. two separate transformers, one to serve normal plant loads and the other for the large fluctuating loads, or 3. a transformer whose kva rating is chosen to meet normal plant load requirements in combination with a series capacitor in the utility feeder circuit ahead of the main transformer. It was found to be economical for the General Electric Company to furnish the main substation as a package unit, Fig. 1, and to accept power from the utility at 34.5 kv.

One factor influencing the design of the plant's primary distribution system was the existence of several new and complete load center unit substations which had been previously ordered and manufactured for use in another proposed new General Electric plant. Since the plans for the other plant had been deferred after
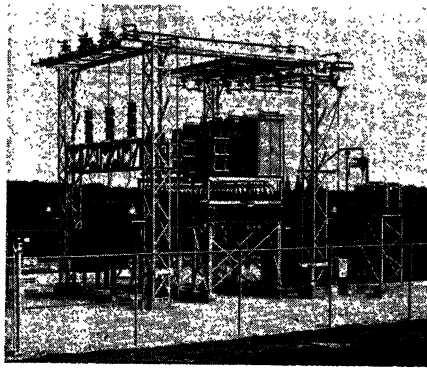
Fig. 1. Outdoor package substation, rated 5,000 kva, 34.5 to 13.2 kv with series capacitor

the switchgear was manufactured, this equipment was utilized for the General Electric Bloomington plant to meet the very short schedule which had been established for plant completion and operation. The primary-voltage rating of these substations was 13.2 kv, and the low-voltage rating 480Y/277 volts. The 13.2-kv rating of the load center unit substation transformers had been chosen to match the utility voltage available at the other General Electric plant site. Thus, the selection of voltage for distribution was dictated by the existing equipment, namely, 13.2 kv. From an economic standpoint, 4.16-kv plant distribution might have been a better choice according to present-day practices. Special considerations in this case, however, precluded selecting the distribution voltage strictly on an economic basis. The 480Y/277 volt secondary distribution is in accordance with modern practice and would have been chosen regardless of primary-voltage level. The 13.2-kv system is operated with the transformer

neutral grounded through a 300-ampere resistor. This value of ground-fault current assures proper operation of the transformer differential and ground-fault relaying schemes.

The over-all distribution system is a simple radial type, Fig. 2 with provision for future tie cables between various load center unit substations to provide a secondary selective system should it seem advisable at a later date. The 480-volt tie cables are not installed at present, but their future installation will present no problem, as it will be made with interlocked armor cable in existing racks or trays.

The 13.2-kv power is brought underground from the main substation which is some 500 feet from the main building. The underground cable is lead-covered varnished-cambric-insulated, and is spliced into the interlocked armor as it comes up into the building. All indoor 13.2-kv distribution within the plant is made with interlocked armor cable. The radial system is designed, except for the welding substation, to limit the maximum short-circuit current on the 480-volt system to 25,000 amperes. Thus, motor control centers and distribution busway of a standard rating are used throughout the plant. The high-voltage side of each substation is furnished with an air-interrupter-type switch, with the switch key interlocked with the secondary breaker. It is interesting to note that in order to preserve maximum factory floor area for manufacturing purposes, the substations, Fig. 3, are located on balconies. In addition to saving space, this arrangement locates the substations where they are isolated and thus free from tampering by unauthorized persons.

Secondary Distribution, 480 Volts or Less

The loads to be served by this power system can be segregated into three main categories: 1. General lighting and auxiliary motor load, 2. Relatively heavy
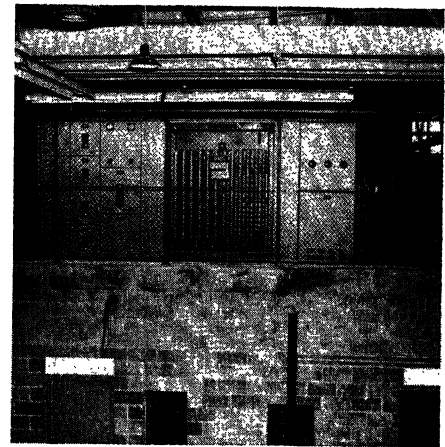
Fig. 3. Load center unit substation located on balcony and enclosed with steel fence

welding loads, 3. Heavy testing loads of low power factor.

The method of power distribution employed to serve each load category is tailored to satisfy the requirements dictated by the load characteristics. Thus, in the following text each load classification is discussed individually.

LIGHTING AND GENERAL DISTRIBUTION

The general lighting load and the motor load are similar to those found in many factories and present no particular problem. Except for certain isolated areas, fluorescent lighting is used throughout the plant, with the lighting installation at 277 volts. The general level of illumination in the manufacturing area is 40 foot-candles. This intensity is increased to 50 foot-candles in a large section of the plant where a large portion of the work to be performed is the hand assembly of small devices. To obtain the additional illumination, fluorescent fixtures are added at right angles to the main rows of fixtures approximately every 15 to 20 feet; see Fig. 4.

Miscellaneous auxiliary outlets for 110-volt service are located on selected columns throughout the factory area. These outlets are served from several small dry-type transformers located in the truss structure of the roof. In the office area where a greater concentration of 110-volt equipment is found, a network of underfloor-fiber duct and floor outlets is established. Oval duct houses the 110-volt wiring, and also the telephone wiring with the power cable brought into the junction points of the duct network; see Fig. 5. The source of power for the office area is three single-phase 25-kva dry-type transformers stepping down from 480 volts and located in service areas of the office extension.
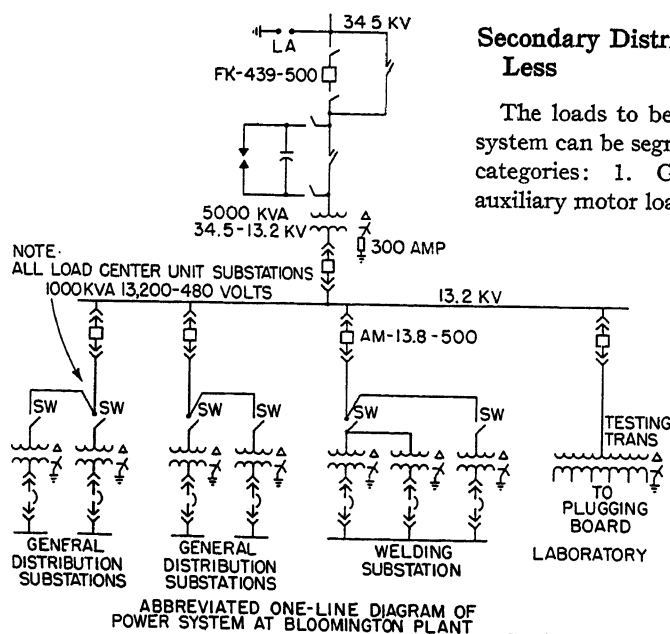
Fig. 2. Abbreviated 1-line diagram of power system

FIG. 2

ABBREVIATED ONE-LINE DIAGRAM OF POWER SYSTEM AT BLOOMINGTON PLANT

Fig. 4. High concentration of fluorescent lamp fixtures in assembly area showing 480-volt maintenance receptacle on column in foreground
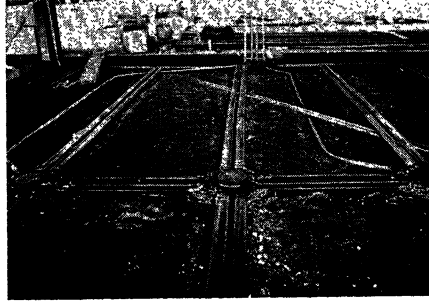
Fig. 5 (above). Oval duct installation in office area prior to pouring concrete

Fig. 6 (right). Close-up of 480-volt receptacle and disconnecting switch

The general 480-volt distribution throughout the factory area is by 400-ampere plug-in busway connected to the substations by interlocked armor cable. A 4-conductor bus is used with the fourth bar reserved for grounding the frames of the machine tools and other machinery. During normal operation this conductor carries no current. In general, taps from the busway are made by means of fused safety switches with standard flexible bus-drop cables connecting the machine to the fused switches. This arrangement provides a completely flexible layout, allow-ing for partial or complete rearrangement of the plant utilization equipment with minimum disturbance to the plant power system layout.

One interesting innovation appearing at this modern plant is a network of polarized plug-in receptacles for auxiliary 480-volt power. Some of the portable maintenance devices require voltages above 110 volts, and it was decided that these devices could best be served from the 480-volt outlets, with individual step-down transformers mounted on the maintenance equipment to give the proper utilization voltage. The polarized plug is mounted on a column, Fig. 4, with its disconnecting switch, Fig. 6, mounted "within sight" of the maintenance crew. This circuit is tapped from the overhead plug-in busway through a fused safety switch.

WELDING LOADS

Metal enclosures must be fabricated for most of the devices which are manu-factured in the plant, resulting in a rela-tively heavy welding load. The welders vary in size from 4 to 400 kva, and for the most part are well concentrated in two closely located areas of the plant. A few small, remotely located welders are fed directly from the general 480-volt distribution system. The heavy concen-tration of welders is supplied from a 3,000-kva welding substation located near the fabricating area. The welding substation is actually made up of three 1,000-kva transformers feeding a common bus. The use of three smaller transformers re-quires more space than would a single 3,000-kva unit, but since the major components of the 1,000-kva units were available, this arrangement was utilized.

Distribution from the welding substa-tion to the two welding areas is made by

FLICKER OF INCANDESCENT LAMPS CAUSED BY RECURRENT VOLTAGE DIPS
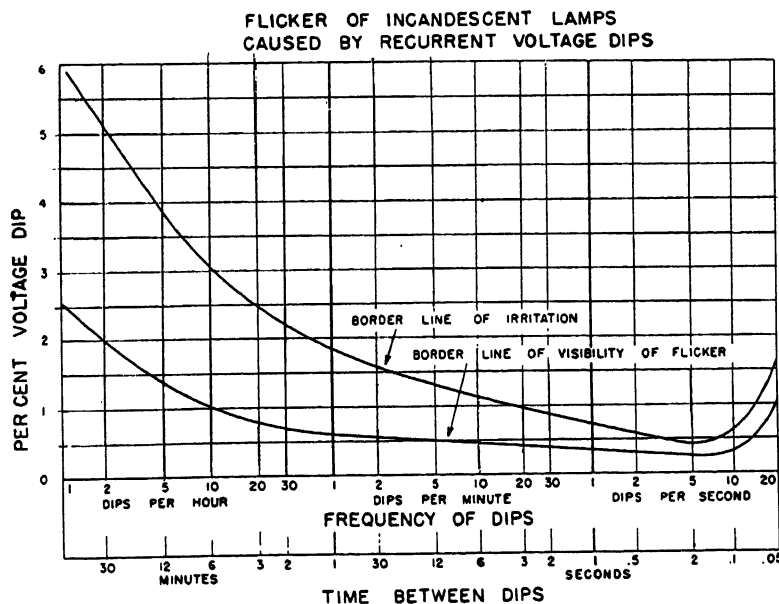
Fig. 7. Relation of voltage fluctuations to their frequency of occurrence: incandescent lamps
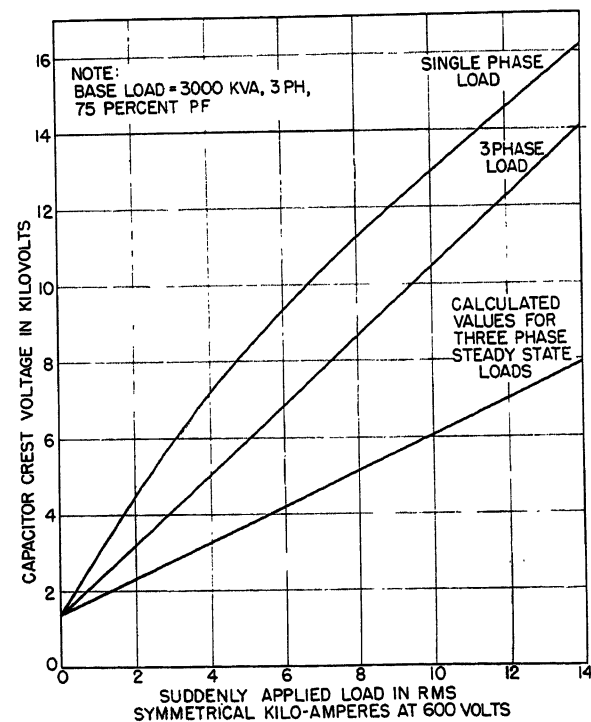
Fig. 8. Relationship of voltage across series capacitor and suddenly applied load
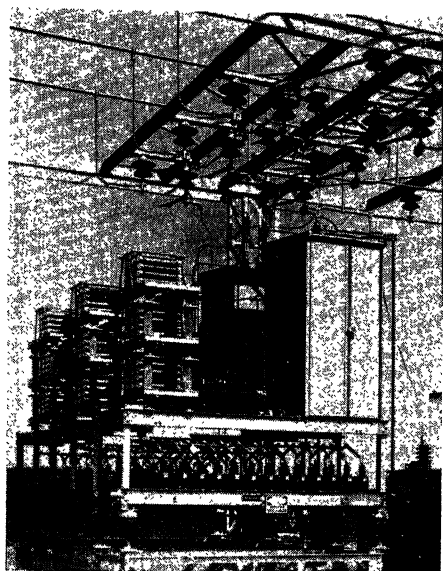
Fig. 9. Close-up of series capacitor installation at package substation

a low-impedance type of busway in order that voltage drop be held to a minimum. This results in a neat and compact arrangement for distribution to loads located as they are in this section of the plant. Each of two bus runs is fed from an *AK-1-50* circuit breaker at the substation. The welding substation is located on a balcony and fed directly from its own breaker at the main outdoor substation.

TESTING LOADS

The testing loads are found in the development laboratory, where the tests performed are many and varied. Power must be available for the development testing of equipment to be used not only under present-day codes and standards, but also under standards expected to exist in the future. The magnitude of such loads ranges from simple control circuit service to that of load life testing of control devices and even to limited short-circuit tests. Load life testing of devices involves continuously pulsating loads. The loads are applied for only a fraction of a second and then are off for a few seconds, with the cycle continuously repeated for a specified period. The current magnitude of these tests ranges as high as the locked rotor current of a 200-horsepower 550-volt 3-phase induction motor. Three such tests may be operating at one time but will be staggered by sequence control so they do not "peak" at the same time. Short-circuit tests are of very short duration and are performed only occasionally. Normally, the magnitude of current required for such tests is specified in the codes of the Underwriters Laboratory. The values of current considered may be as high as several thousand amperes on a 600-volt

3-phase circuit. The remaining test requirements range between the frequency of the load life test and the maximum power of the controlled short-circuit test. Any of these tests may be performed over a voltage range of approximately 120 to 720 volts. Other experimental and development work may require even slightly higher voltages and, in addition, the voltage must be adjustable in small steps over the complete range.

It is obvious that a special low-impedance testing transformer would be needed to meet the aforementioned requirements. Load ratio control on the test transformer with a range of plus or minus 16 per cent, and with appropriate taps on the low-voltage winding, gives the desired voltage steps over the complete range and allows for adjustment underload. The operation of the load ratio control is supervised by the laboratory personnel performing the test, with a control switch and position indicator located at the test control panel. It is evident from the foregoing that because of the nature of the tests, the load consists of quite high peaks, but the thermal load on the transformer is fairly small. Thus the laboratory transformer is designed specifically to meet these requirements; no other demands are required of it. The transformer is oil-filled and is located outdoors just on the other side of the building wall from the testing area. Low-voltage entrance into the building is by cable entering through a duct and connected directly to a distribution and tap selector board. By necessity, the length of the incoming leads must be as short as possible in order to hold cable impedance to a minimum. Because of the nature of the duty imposed on the 13.2-kv circuit feeding the test transformer, no other loads are connected to it and it is served from its individual circuit breaker at the main substation. Remote control of the feeder breaker is available at the control board in the laboratory.

The special transformer is fed from the same 13.2-kv bus as the load center substations in the manufacturing and office areas. The effect of the peaking loads resulting from welding requirements and development testing would therefore be voltage dips on the plant bus unless some corrective measures are taken. Since the power peaks occur frequently during certain tests, a minor voltage dip is very noticeable.

Series Capacitor Application

GENERAL CONSIDERATIONS

Determining the most satisfactory
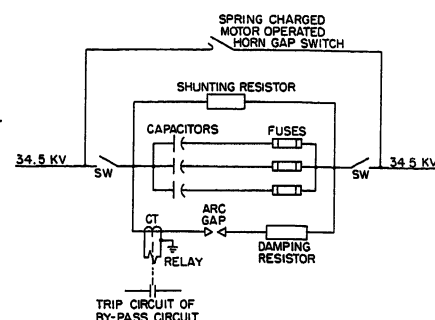
Fig. 10. Schematic diagram of series capacitor and protective scheme

method of compensating for the frequent voltage dips resulted in a comprehensive economic and engineering analysis and presented something of a challenge in the planning of the power system. As has been mentioned in the opening paragraphs, the results of such studies indicated the existence of a definite advantage in the application of a series capacitor to the system.

It is practical in most industrial applications to regulate voltage by either automatic load ratio control on the transformer or by applying an induction voltage regulator in the feeder circuits. It must be remembered, however, that the testing and welding loads consist of high peaks of very short duration. The peak lasts only a small fraction of a second and voltage regulating equipment as normally applied will not act in such a short time. By contrast, a series capacitor offers its voltage correction almost instantaneously. In usual industrial circuits the time constant of voltage recovery following sudden application of load is in the order of 0.01 second for the capacitor. Because of this instantaneous and automatic response of the series capacitor, it is admirably suited to compensate for voltage drop or fluctuation associated with intermittent, fluctuating, or suddenly applied load as is found here.

The series capacitor can be considered as a negative reactance and thus its insertion into the circuit compensates for the reactances of the main transformer and the plant primary feeder. To the limit of compensation desired the capacitor thus shows a voltage rise equal to the voltage drop in the transformer and feeder and tends to hold a more stable voltage pattern on the feeder bus.

CAPACITOR RATING

An economic evaluation was made to determine the amount of correction which should be made with the series capacitor. Economics indicated that the limit should be such that when a low power factor

testing load of 2,000-kva peak was imposed on the normal plant load, the voltage drop on the main 13.2-kv bus will not exceed 0.5 per cent. It is apparent, Fig. 7, that with this voltage drop the frequency of application of a peak load could be seven times per minute without causing any visible flicker in incandescent lamps. The frequency could be increased considerably before the flicker actually becomes annoying (three per second). This allows some margin for multiple peaks occurring when several life tests are running simultaneously. It will be noted from Fig. 7 that the above values apply for incandescent-type lighting. Voltage dips of greater magnitude can be tolerated where fluorescent-type lighting is applied. Although the major portion of the lighting in this plant is of the fluorescent type consideration must be given to incandescent lighting for display floodlights and work-lights on machine tools.

Economic evaluation shows the most favorable location of the series capacitor to be the 34.5-kv side of the transformer and the aforementioned limitations result in a capacitive reactance requirement of 19 ohms. The required voltage rating of the capacitor is, of course, determined by the current which must flow through it, in combination with its reactance. Consideration must be given to not only steady-state values but transient peak voltages.

Tests performed on the transient analyzer determined the transient voltages which would appear when various values of peak loads are imposed on the normal plant load. The results of the study are plotted in Fig. 8. From these data the proper voltage rating was chosen to comply with the testing program established by the laboratory personnel.

The capacitors and the associated protective equipment are not insulated for the full line voltage of the system, since this would result in large and uneconomical equipment. They are located on a common steel rack, Fig. 9, which is mounted on pedestal-type insulators of the proper voltage level.

CAPACITOR PROTECTIVE EQUIPMENT

It was uneconomical in this installation to apply capacitors suitable to handle the magnitude of current which would flow during certain high-current tests, since

these tests are run very infrequently. Operating procedures are thus established in the laboratory to dictate that these tests be run only at certain hours and that the series capacitor be removed from the circuit at the time. This would be done with a motor-operated by-passing switch. Conditions may exist, however, where proper procedures are not followed and it is thus necessary to provide protective equipment which will automatically protect the capacitor.

This protective equipment utilizes an arc gap as the basic protection. The arc gap setting is such that when voltage across the capacitor reaches approximately twice its rated value the arc gap sparks over to prevent damage to the capacitor. The gap in turn is relieved of continuous duty by a short-circuiting device.

Available standard protective equipment utilizes a contactor to by-pass the arc gap, with the voltage for the contactor operating coil established by the arc current flowing through a damping resistor. The current flow must be great enough to establish the necessary operating voltage and must persist for a length of time sufficient to allow for operation of the contactor. Thus a high momentary current, lasting for only a few cycles, will flash over the arc gap but may be of insufficient duration to operate the contactor. The characteristics of an arc gap are such that once the gap has flashed over, the arc may be maintained by a lower magnitude of current than was required to establish it. The voltage developed across the dropping resistor may be insufficient with this lower magnitude of current to operate the contactor. This would result in continuous current flow across the arc gap and lead to eventual destruction of the electrodes.

From the foregoing it is obvious that the standard protective equipment, may not perform satisfactorily with the high-magnitude short-duration type loads imposed on the series capacitor at the Bloomington plant. Thus the protective scheme was revised for this installation.

The revised protective scheme employs an instantaneous relay which would pick up whenever the arc gap flashed over. The relay acts much faster than a contactor and thus will operate during the time the peaking load is im-

posed on the system. Current transformers in the gap circuit isolate the relay from the 34.5-kv system. A spring-charged, motor-operated closing mechanism is applied to the horn-gap switch normally used to by-pass the capacitor manually. The contacts of the instantaneous relays are wired in the tripping circuit of the spring-charged mechanism; see Fig. 10. Whenever the arc gap spills over, the switch is closed automatically, thus relieving the electrodes of continuously carrying current.

To summarize the operation of the series capacitor as utilized in this particular system, the following conclusions are drawn:

1. For normally scheduled routine tests in the laboratory the series capacitor will remain in the circuit.

2. If the capacitor is not intentionally removed from the circuit when irregular tests are run, it will automatically be by-passed by protective equipment.

3. When a fault occurs within certain zones of the primary system (which would also result in a current flow in excess of the capacitor rating), the protective equipment will supervise a function which will by-pass the capacitor and, if required, trip the main circuit breaker.

Conclusion

The general load requirements of this plant are typical of these found in new industrial installations. Not so commonplace, however, are the specialized requirements at the development laboratory and the compensatory measures taken to reduce their effects on the balance of the plant power system.

By combinations of properly applied equipment as outlined in this paper, the power requirements of both the manufacturing and testing facilities are fulfilled. The application involved a number of technical and economic comparisons. The results of these comparisons were examined and integrated with the considerations which must be given to such items as existing utilization or distribution equipment, normal and abnormal load requirements, and the practices and capabilities of the local utility. From the foregoing evolved the plans for the power system to fulfill not only present needs but expected future requirements.

Short-Time Memory Devices in Closed-Loop System—Steady-State Response

T. W. SZE
ASSOCIATE MEMBER AIEE

J. F. CALVERT
FELLOW AIEE

POTENTIAL uses of short-time memory devices in linear, closed-loop systems are shown, and the primary concern is with follower-type systems. In so far as possible the proposed AIEE symbols will be employed.[1]

Two earlier papers[2,3] presented, in rather general terms, the possible uses of these devices as compensators, or lead networks, for open-loop systems. In these it was demonstrated that a marked increase in bandwidth could be achieved for a given control system when there was cascaded with it a compensator which incorporated an appropriately designed short-time memory unit. However, these arrangements introduced two matters of concern: 1. the control system might be linear but its parameters not known with accuracy sufficient to permit designing a fully adequate compensator, or 2. the control system might contain nonlinearities which cast serious doubt on the likelihood of obtaining satisfactory performance in any open-loop system predicated on linear theory.

These potential limitations seemed to suggest a feedback system but, since discrete time delays were a feature of the short-time memory devices and, hence, would be literally built into the closed loop, trouble with steady-state stability and difficulty with the computation of transient performance both seemed likely. It might be argued, of course, that, if the compensator is moderately well designed, the percentage of the signal which is fed back need not be large, only large enough to provide a sort of "fine" control.

In the following it will be shown that short-time memory devices may be employed to advantage in feedback and also in feedforward-feedback systems. This paper is restricted to steady-state performance.

Nomenclature

TIME-ORIENTED QUANTITIES

f = frequency of the command function
$s = re^{j\theta}$ where r and θ are magnitude and angle
  = complex variable of Laplace transformation calculus
t = instantaneous time
T = a period of time in units of t
T_d = total time delay possible within the short-time memory unit
$\omega = 2\pi f$ = angular frequency
μ = a real or complex root of $D(s) = 0$

VARIABLES WHICH ARE FUNCTIONS OF TIME

v or $v(t)$ = command function (or input)
q or $q(t)$ = either controlled or indirectly controlled variable (output or response)

TRANSFER FUNCTIONS

$\dfrac{N(s)}{D(s)}$ = transfer function of the control system where both $N(s)$ and $D(s)$ are polynomials of s

$\dfrac{D_x(s)}{N_x(s)}$ = transfer function of the compensator where $D_x(s)$ is a transcendential, as shown in equation (1), and $N_x(s)$ is a polynomial designed to approximate $N(s)$

In these terms s may be replaced by $j\omega$ to give the steady-state sinusoidal transfer functions. Additional subscripts as $N_1(s)$ or D_{XJ}, etc., may be introduced and here reference should be made to the associated figures.

$G = G(s) = N(s)/D(s)$
$B = B(s) = D_x(s)/N_x(s) = D_{BX}(s)/N_{BX}(s)$
$J = J(s) = D_{JX}(s)/N_{xX}(s)$
$A(\omega)$, $B(\omega)$ = factors in $D_x(j\omega)$ as defined in equation (5)
$J(s), f(s), X(\omega), Y(\omega), U(\omega), V(\omega), R$ = functions defined where used
ψ = the argument of $f(s)$

PARAMETERS

a_n, B_k, K, b_m = real numbers defined in part by usage in the text. Here m, n, and k take no various values
C = a closed path along which integration is carried out
M = a real positive number used to signify the magnitude of $F(j\omega)$

SUBSCRIPTS

$1, 2, \ldots, k, l, m, n$ = positive integers used as subscripts. In one instance, a and b are used as subscripts to indicate the start and finish of a line integral

Short-Time Memory Units as Compensators

The two developments, referred to in the foregoing, concerning open-loop systems with short-time memory units serving as part of the compensator devices were based on different performance criteria and the presentations exhibited different mathematical approaches. Compensators designed on either basis (or a combination) can be employed for the closed-loop systems to be described herein. Therefore, as background, these earlier developments will be contrasted, and this is done in a mathematical form most suitable for the closed-loop development to follow.

First, a few general statements concerning the open-loop system of Fig. 1

$$q = D_x(s)\left(\frac{1}{N_x(s)}\right)\left(\frac{N(s)}{D(s)}\right)v \qquad (1)$$

where

$$N(s) = b_0 + b_1 s + b_2 s^2 + \ldots + b_m s^m$$

$$D(s) = a_0 + a_1 s + a_2 s^2 + \ldots + a_n s^n$$

$$N_x(s) = b_0' + b_1' s + b_2' s^2 + \ldots + b_m' s^m$$

$$D_x(s) = B_0 + B_1 \epsilon^{-T_1 s} + B_2 \epsilon^{-T_2 s} + \ldots + B_k \epsilon^{-T_k s}$$

and in which

$$b_0' \approx b_0, \ b_1' \approx b_1 \ldots b_m' \approx b_m, \text{ and } k \geq n > m$$

and where these and the parameters B_0, B_1, $B_2 \ldots B_k$ are all finite real numbers.

Assumptions and facts which, subsequently, are employed concerning these functions are the following:

1. The individual functions $N(s)$, $D(s)$, and $N_x(s)$ are polynomials. They are rational and analytic in the finite domain of s. They contain no poles in this region and it is assumed that they arise from physical systems of such form that, as individual functions, they contain no zeros in the right half of the s plane.

2. The term $N_x(s) \approx N(s)$. In fact, it is assumed that a linear passive network can be designed such that its transfer function is $[1/N_x(s)]$.

3. The control system itself described by $[N(s)/D(s)]$ is assumed as given and unalterable.

4. The transcendental function $D_x(s)$ is analytic in the finite s domain. It possesses in this area no poles and, in any finite area, at most, a finite number of zeros; it will be demonstrated later under one criterion that for the values of s of primary concern, designs can be established such that $D_x(s) \approx D(s)$.

The foregoing describes the open-loop system in a single mathematical form. Two sets of criteria and the resulting design equations are discussed next.
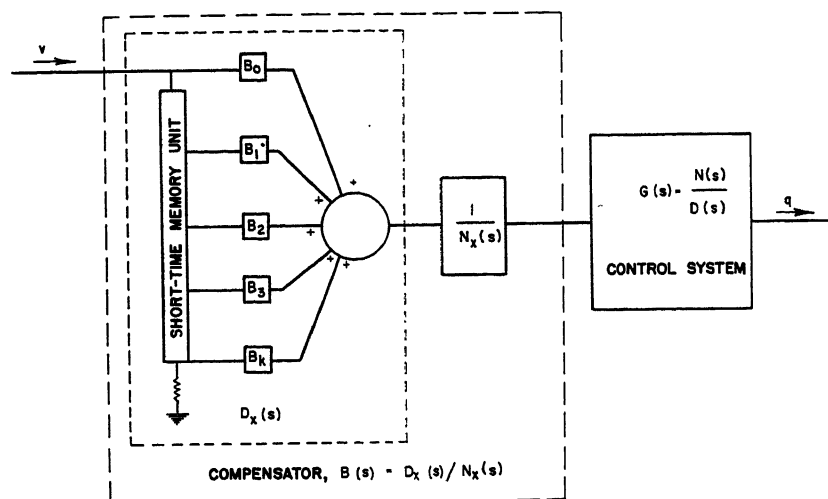
Fig. 1. Open-loop system with compensator which incorporates short-time memory unit

B (s) = 4.88−7.16 $\epsilon^{-\frac{1}{2}s}$ + 3.28 ϵ^{-s}

Angular Frequency, ω , Radians Second

A — Uncompensated system, G = $\frac{1}{1+.8s+s^2}$

B — Compensated by a 3-tap SCC compensator, $T_d = 1$

Fig. 3. Open-loop system performance based on criterion associated with a step command function

Polynomial Command Function

In this approach it is assumed that the command function v was representable, at least over successive short periods, by a polynomial in time. The criterion selected concerning the output q (which may be either the directly or the indirectly controlled variable) is that, after action is initiated, q can be completely defined by the following components:

1. At every instant of time, q will contain a component which is identical with the polynomial v.

2. For all time after a specified period, which is usually taken equal to or less than one half the longest natural period of oscillation of the control system, all periodic errors will reduce to and remain at zero;

3. Within the same period of time, if v contained more than just a constant term, all aperiodic errors will reduce to and remain at zero.

These criteria lead to the design equations for $D_x(s)$. However, in the illustration to follow the polynomial input

will be reduced to its simplest form, i.e., a step input and, in consequence, no aperiodic error is to be encountered

$$D(s) = a_n(s-\mu_1)(s-\mu_2)\ldots(s-\mu_n) \qquad (2)$$

where $\mu_1, \mu_2 \ldots \mu_n$ are the roots of $D(s) = 0$.

From equation 1, letting $k=n$

$$D_x(s) = B_0 + B_1\epsilon^{-T_1s} + B_2\epsilon^{-T_2s} + \ldots + B_n\epsilon^{-T_ns} \qquad (3)$$

To satisfy the second criterion stated in the foregoing, the following relations are introduced

$$B_0 + B_1\epsilon^{-\mu_1 T_1} + B_2\epsilon^{-\mu_1 T_2} + \ldots + B_n\epsilon^{-\mu_1 T_n} = 0$$
$$B_0 + B_1\epsilon^{-\mu_2 T_1} + B_2\epsilon^{-\mu_2 T_2} + \ldots + B_n\epsilon^{-\mu_2 T_n} = 0$$
$$B_0 + B_1\epsilon^{-\mu_n T_1} + B_2\epsilon^{-\mu_n T_2} + \ldots + B_n\epsilon^{-\mu_n T_n} = 0 \qquad (4A)$$

To satisfy the first criterion, the final value theorem in Laplace transformation calculus is employed to state that, for a step input

$$1 = \lim_{s \to 0} \frac{s(B_0 + B_1\epsilon^{-T_1s} + \ldots + B_d\epsilon^{-T_ns})}{s(a_0 + a_1s + a_2s^2 + \ldots + a_ns^n)}$$

From which

$$B_0 + B_1 + B_2 + \ldots + B_n = a_0 \qquad (4B)$$

For the case of a step input, equations 4(A) and 4(B) form the design equations in which $T_1, T_2 \ldots T_n$ are to be selected and then $B_0, B_1, B_2 \ldots B_n$ are computed. These are the design equations for the

E $^{-jT_d\omega/2}$ $\left[B_{a_0} + B_{a\ell}\cos T_\ell\omega + jB_{b\ell}\sin T_\ell\omega \right]$ v

Fig. 2 (left). Development of $D_x(j\omega)$ for sinusoidal command function

Fig. 4 (right). Open-loop system performance based on criterion associated with sinusoidal command function

3 - Tap Compensator = B (s) = 4.8−7.6 $\epsilon^{-\frac{1}{2}s}$ + 3.2 ϵ^{-s}

5-Tap Compensator = B(s) = −1.6 + 23.56 $\epsilon^{-\frac{1}{4}s}$−39 $\epsilon^{-\frac{1}{2}s}$ +19.26 $\epsilon^{-\frac{3}{4}s}$ +1.62 ϵ^{-s}

Angular Frequency, ω , Radians / Second

A — Uncompensated system, G = $\frac{1}{1+.8s+s^2}$

B — Compensated by a 3-tap TSC compensator, $T_d = 1$

C — Compensated by a 5-tap TSC compensator, $T_d = 1$

short-time-memory unit for the step input and the criteria set forth in the foregoing.

Sinusoidal Command Function

Here, it is assumed that the command function v is sinusoidal. This time the criteria for the controlled variable q are concerned only with the steady-state response as follows:

1. The amplitude of the controlled variable q when plotted versus frequency will be of a desired shape and, usually this is taken to be nearly flat over a wider frequency band than was the case for the control system alone, regardless of the design of the latter.

2. Some angular lag will be accepted for q with respect to v. Usually, the design will be such that a linear phase lag results because, for a follower system, this means a delayed output though not distorted in the process.

To satisfy the foregoing criteria, the expression for $D(j\omega)$ is established in polynomial form. Then $D_x(j\omega)$ is written as the product of two terms. The first provides only the linear phase lag. The second provides a function which will be made to match nearly $D(j\omega)$ over as wide a frequency band as seems practicable in terms of saturation and other limitations of real equipment. Details of the foregoing are

$$D(j\omega) = a(\omega) + jb(\omega) \qquad (5)$$

where

$$a(\omega) = a_0 + a_2\omega^2 + a_4\omega^4 + \ldots + (-1)^{\frac{1}{2}n}a_n\omega^n$$

$$b(\omega) = a_1\omega + a_3\omega^3 + a_5\omega^5 + \ldots + (-1)^{\frac{1}{2}n-1}a_{n-1}\omega^{n-1}$$

assuming that n is an even number.

Reference to Fig. 2 shows that

$$D_x(j\omega) = \epsilon^{-j\frac{1}{2}T_d\omega}[A(\omega) + jB(\omega)] \qquad (6)$$

where T_d is the total delay time of the memory unit and

$$A(\omega) = B_{a0} + B_{a1}\cos T_1\omega + B_{a2}\cos T_2\omega + \ldots + B_{ak}\cos T_k\omega$$

$$B(\omega) = B_{b1}\sin T_1\omega + B_{b2}\sin T_2\omega + \ldots + B_{bk}\sin T_k\omega$$

It is desired to make $A(\omega) \approx a(\omega)$, and $B(\omega) \approx b(\omega)$.

These last requirements can be given physical representation in one of several ways. However, it is extremely important to keep $1/2T_d$ small. A good representation with small time delay was accomplished successfully only with a modified Taylor's series type of compensator (TSC). For this design the trignometric terms of equation 6 are expanded in series forms, the coefficients of the appropriate powers of ω are matched to those of

equation 5, and then a certain number of additional coefficients of ω in equation 6 are summed to zero.

These procedures yield equations 7 which are the design equations for the short-time memory unit when a sinusoidal input is assumed and the second set of criteria are employed. In the use of equation 7, the values of the T are chosen and then the values of the B are computed. It will be observed that k is an even integer and greater than n.

$$B_{a0} + B_{a1} + B_{a2} + \ldots + B_{ak} = a_0$$

$$T_1{}^2 B_{a1} + T_2{}^2 B_{a2} + \ldots + T_k{}^2 B_{ak} = (2!)a_2$$

$$T_1{}^n B_{a1} + T_2{}^n B_{a2} + \ldots + T_k{}^n B_{ak} = (n!)a_n$$

$$T_1{}^{n+2} B_{a1} + T_2{}^{n+2} B_{a2} + \ldots + T_k{}^{n+2} B_{ak} = 0$$

$$T_1{}^k B_{a1} + T_2{}^k B_{a2} + \ldots + T_k{}^k B_{ak} = 0 \qquad (7)$$

and

$$T_1 B_{b1} + T_2 B_{b2} + \ldots + T_k B_{b(k-1)} = (1!)a_1$$

$$T_1{}^3 B_{b1} + T_2{}^3 B_{b2} + \ldots + T_k{}^3 B_{b(k-1)} = (3!)a_3$$

$$T_1{}^{n-1} B_{b1} + T_2{}^{n-1} B_{b2} + \ldots + T_k{}^{n-1} B_{b(k-1)} = (n-1)!a_{n-1}$$

$$T_1{}^{n+1} B_{b1} + T_2{}^{n+1} B_{b2} + \ldots + T_k{}^{n+1} B_{b(k-1)} = 0$$

$$T_1{}^{k-1} B_{c1} + T_2{}^{k-1} B_{b2}{}^c + \ldots + T_k{}^{k-1} B_{b(k-1)} = 0$$

Figs. 3 and 4 show calculated performance curves for open-loop systems based respectively on the first and on the second design criteria discussed in the foregoing.

Short-Time Memory Devices in Feedback Systems

STABILITY

Figs. 5(A) and (B) show two systems having the transfer function $F(s)$ where

$$q = [F(s)]v$$

$$F(s) = (1+K)\frac{BG}{1+KBG}$$

$$F(s) = (1+K)\frac{D_x(s)N(s)}{N_x(s)D(s)\left(1 + K\dfrac{D_x(s)N(s)}{N_x(s)D(s)}\right)} \qquad (8)$$

where the $(1+K)$ factor compensates for the loss of gain and

$$G = G(s) = N(s)/D(s) \qquad (9)$$

$$B = B(s) = D_x(s)/N_x(s)$$

All terms on the right side of equations 9 are as defined for equation 1; and K is a positive real number which, as will be shown later, is usually chosen to be less than unity.

Routh's and Hurwitz's criteria[4] deal with the coefficients of polynomials. Hence, because of the presence of the transcendental expression $D_x(s)$ in the denominator of equation 8, it appears that the Nyquist's criterion is the one to employ in a stability study. The expression $D_x(s)$ and $N_x(s)$ are not seen in the more conventional feedback system, but will be employed both in the feedback and in the feedforward-feedback systems to be discussed in the following. Therefore, it seems advisable to carry through a brief development of the Nyquist criterion for the specific systems at hand.

Use is made of the assumptions accompanying equation 1. It is seen in the finite domain of s that $D_x(s)N(s)$ and $N_x(s)D(s)$ provide no poles; and, in accord with the assumptions, provide no zeros in the right half of the s-plane. Therefore, it is desired to know only whether or not any zeros are provided in the right-half plane by the function

$$f(s) = 1 + K\frac{D_x(s)N(s)}{N_x(s)D(s)} \qquad (10)$$

Since each term on the right is analytic and possesses the characteristics first cited, Cauchy's integral theorem[5] may be modified to state: The integral $(1/2\pi j)\cdot\int_c[f'(s)/f(s)]ds$ taken in a positive sense around the boundary c of the right-half plane is equal to the number of zero points of $f(s)$ lying within the enclosed domain, each being counted a number of times equal to its order.

In general, between any two points s_a and s_b

$$\frac{1}{2\pi j}\int_{s_a}^{s_b}[f'(s)/f(s)]ds = -\frac{j}{2\pi}\ln\frac{|f(s_b)|}{|f(s_a)|} + \frac{1}{2\pi}(\psi_a - \psi_b) \qquad (11)$$

When the integral is taken around a simple closed path $|f(s_b)| = |f(s_a)|$, the first term on the right side of equation 11 becomes zero and the second is equal to the integral number of revolutions of $f(s)$ made in traversing the closed path C.

Consider a D-shaped closed path made up of the imaginary axis plus a semicircular path lying in the right-half plane with the center at the origin. Let $s = re^{j\theta}$. It is noted, then

$$\lim_{r\to\infty} D_x(s) \leqq |B_0| + |B_1| + |B_2| + \ldots + |B_k| = \text{a finite real value} \qquad (12)$$

$$\lim_{r\to\infty}\left|\frac{N(s)}{N_x(s)}\right| \lim_{r\to\infty}\left|\frac{\sum\limits_{l=0}^{l=m} b_l r^l e^{jl\theta}}{\sum\limits_{l=0}^{l=m} b_l' r^l e^{jl\theta}}\right| = \left|\frac{b_m}{b_m'}\right| \approx +1 \qquad (13)$$
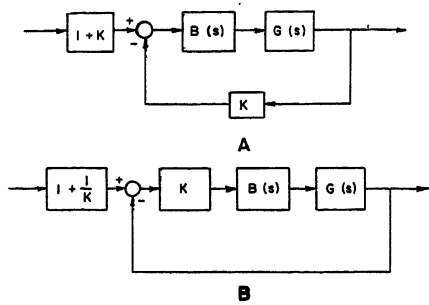
Fig. 5. Feedback scheme using loop gain to obtain stabilizing system or improved performance

A—Loop gain K in feedback path
B—Loop gain K in forward path

$$\operatorname*{Lim}_{r\to\infty}\ |D(s)| = \operatorname*{Lim}_{r\to\infty}\left|\sum_{l=0}^{l=n} a_l r^l e^{jl\theta}\right| = +\infty \tag{14}$$

Considering the semicircular path, there is some finite value R such that for $r > R$ and $1/2\pi \geqq \theta \geqq -1/2\pi$

$$\left|\frac{D_x(s)N(s)}{N_x(s)D(s)}\right| < 1$$

then

$$2 > |f(s)| > 0$$

Hence, for $r > R$, the function $f(s)$ has no poles or zeros on the imaginary axis or in the right-half plane. Therefore, as r increases beyond R, the path of integration for $\int_c [f'(s)/f(s)]ds$ will go through no singular points.

Furthermore

$$f(re^{j\theta}) = 1 + A_m e^{j\rho}$$

Fig. 7. Frequency response of system compensated by feedback using memory units in the forward path

where

$$A_m e^{j\rho} = \frac{D_x(re^{j\theta})N(re^{j\theta})}{N_x(re^{j\theta})D(re^{j\theta})}$$

As r approaches infinity

$$\operatorname*{Lim}_{r\to\infty} A_m = 0$$

and

$$\operatorname*{Lim}_{r\to\infty} f(s) = 1 + (0 + j0) = 1 \tag{15}$$

In consequence, along the semicircular path as r approaches infinity, $(\psi_b - \psi_a)$ approaches zero, and nothing is contributed to the total integral around the right-half plane of s.

Fig. 6 (left). Design of loop gain K in the feedback path

Fig. 8 (below). Feed-forward-feedback systems

Therefore, just as in the systems where $N_x(s)$ and $D_x(s)$ were not employed, it is possible to write for the present system

$$\frac{1}{2\pi j}\int_c \left(\frac{f'(s)}{f(s)}\right)ds$$

$$= \frac{1}{2\pi j}\int_{j\omega=-j\infty}^{j\omega=+j\infty} \left(\frac{f'(j\omega)}{f(j\omega)}\right)d(j\omega) \tag{16}$$

and Nyquist's criterion follows showing that by plotting

$$[f(j\omega) - 1] = K\frac{D_x(j\omega)N(j\omega)}{N_x(j\omega)D(j\omega)} = U(\omega) + jV(\omega) \tag{17}$$

on the U-V plane, the sum of zeros of $f(s)$ lying in the right-half plane, each counted a number of times equal to its order, is equal to the number of encirclements of the $-1 + j0$ point on the U-V plane. No encirclements indicates that $f(s)$ of equation 8 represents a stable system.

Example 1.
$$G(s) = \frac{1}{1 + 0.8s + s^2}$$
$$B(s) = 4.8 - 7e^{-.5s} + 3.2e^{-s}$$
Choose $K = 1/3$

Example 2.
$$G(s) = \frac{1}{(1 + \frac{1}{4}s)(1 + \frac{1}{2}s)(1 + s)}$$
$$B(s) = 1.75 - 0.75e^{-s}$$
Choose $K = 0.25$

BG locus of Example 1.
BG locus of Example 2.

(A)

(B)

(C)

(D)

In the foregoing, stability was discussed. Now the performance of stable systems will be considered and, to do so, the K-circles or the modified M-circles are introduced. Let $BG = x + jy$, then from equations 8 and 9 for the magnitude of the gain to be unity,

$$|F| = 1 = (1+K)\left|\frac{x+jy}{1+K(x+jy)}\right| \qquad (18)$$

It follows that

$$(x - x_0)^2 + y^2 = R^2$$

where

$$x_0 = \frac{K}{2K+1}, \text{ and } R = \frac{K+1}{2K+1}$$

This system of circles is plotted together with the open-loop locus $B(j\omega)G(j\omega)$. If the open loop contains a delay compensator, the BG locus will, in general, be a spiral winding inward toward the origin. At the low frequencies the locus of BG will match or coincide with one of the K-circles. For instance, as shown in Fig. 6, the open-loop locus matches with the circle of $K = 0.333$. Therefore, 33.3 per cent of the controlled variable is fed back to obtain a flat response. The resulting characteristics are shown in Fig. 7. For actual computations, a somewhat more convenient procedure is to use the $(1/BG)$ plane.

Short-Time Memory Units in Feedforward-Feedback Systems

Figs. 8(A) through (D) show a group of feedforward-feedback systems. The system of Fig. 8(A) was discussed for a specific application by J. R. Moore,[6] except that he employed compensators for J and B which did not involve short-time memory units.

The stability of each of these four systems will be discussed. Starting with Fig. 8(A)

$$q = \frac{JG_1 + BG_2}{1 + HBG_2} v \qquad (19)$$

Here, there are two noninteracting paths, the outputs of which add to pro-

duce the controlled variable q. The open-loop system has the transfer function

$$JG_1 = \frac{D_{JX}(s)}{N_{JX}(s)} \frac{N_1(s)}{D_1(s)} \qquad (20)$$

The individual terms on the right are analytic; they contain no poles and no zeros in the right-half plane, and so JG_1 represents a stable open-loop system.

The closed-loop transfer function is

$$\frac{BG_2}{1+HBG_2} = \frac{1}{N_2(s)D_2(s)} \times$$

$$\frac{D_{BX}(s) N_2(s)}{\left[1 + H(s)\dfrac{D_{BX}(s) N_2(s)}{N_{BX}(s) D_2(s)}\right]} \qquad (21)$$

The individual terms, or functions of s, are assumed to have the same characteristics as those first ascribed to the terms of equation 20. Hence, so far as stability is concerned, only the bracketed term in the denominator needs to be considered. If $H(s) = K$, the whole stability problem for equation 19 becomes identical to that of the feedback system described by equation 9.

For the system shown in Fig. 8(B)

$$q = \frac{(J+B)G_1}{1+HBG_1} v. \qquad (22)$$

Assuming $H(s) = K$ and recognizing that the remainder of the letters designate functions of s as before

$$q = \frac{(N_{BX}D_{JX} + N_{JX}D_{BX})N_1}{N_{JX}N_{BX}D_1\left[1 + K\dfrac{D_{BX}N_1}{N_{BX}D_1}\right]} v \qquad (23)$$

With the same assumptions as were made previously with respect to the functions of s, stability again depends only on the bracketed term in the denominator.

Next, consider Fig. 8(C) with $H(s) = K$

$$q = \frac{JG_1 + BG_2}{1 + H(JG_1 + BG_2)} v \qquad (24)$$

$$= \frac{D_{JX}N_{BX}D_2N_1 + D_{BX}N_{JX}D_1N_2}{N_{JX}D_1N_{BX}D_2\left[1 + K\left(\dfrac{D_{JX}N_1}{N_{JX}D_1} + \dfrac{D_{BX}N_2}{N_{BX}D_2}\right)\right]} v \qquad (25)$$

For stability only the bracketed group of

terms in the denominator are considered. Finally, consider Fig. 8(D), letting $H(s) = K$

$$q = \frac{(J+B)G_1}{1+HBG_1} v$$

$$= \frac{(N_{BX}D_{JX} + N_{JX}D_{BX})N_1}{N_{JX}N_{BX}D_1\left[1 + K\dfrac{D_{BX}N_1}{N_{BX}D_1}\right]} v \qquad (26)$$

Stability depends only on the bracketed term in the denominator.

In these four cases $H(s) = K$, just to reduce the problem to a form discussed previously (equation 9). This was not necessary as it can be seen that $H(s)$ could be used as a variable in s and the same general procedures would follow.

Summary

The paper gives, as background, a single mathematical presentation for open-loop systems using short-time memory devices as compensators when two different sets of criteria are imposed for the design. Stability and performance are presented for a feedback system, and stability is discussed for several simple feedforward-feedback systems employing short-time memory devices in the designs. These may be designed by either of the procedures used with the open-loop systems.

References

1. PROPOSED SYMBOLS AND TERMS FOR FEEDBACK CONTROL SYSTEMS, AIEE Committee Report. *Electrical Engineering*, vol. 70, no. 10, Oct. 1951, pp. 905–09.

2. SIGNAL COMPONENT CONTROL, D. J. Gimpel, J. F. Calvert. *AIEE Transactions*, vol. 71, pt. II, Nov. 1952, pp. 339–43.

3. THE APPLICATION OF SHORT-TIME MEMORY DEVICES TO COMPENSATOR DESIGN, D. J. Ford, J. F. Calvert. *AIEE Transactions*, vol. 73, pt. II, May 1954, pp. 88–93.

4. TRANSIENT IN LINEAR SYSTEMS (book), M. F. Gardner, J. L. Barnes. John Wiley and Sons, Inc., N. Y., vol. I, 1942, pp. 197–201.

5. THEORY OF FUNCTIONS, PART I (book), K. Knapp. Dover Publications, New York, N. Y., 1945, pp. 131–34.

6. COMBINATION OPEN-CYCLE CLOSED-CYCLE CONTROL SYSTEMS, J. R. Moore. *Proceedings, Institute of Radio Engineers*, New York, N. Y., vol. 39, Nov. 1951, pp. 1421–32.

7. A PHASE-PLANE APPROACH TO THE COMPENSATION OF SATURATING SERVOMECHANISM, A. M. Hopkin. *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 631–39.

Feedback in Contouring Control Systems

F. J. ELLERT
ASSOCIATE MEMBER AIEE

A CONTOURING control system is a position control system whose primary function is to position the cutter of a machine tool in accordance with the contours of a template. The General Electric Company's contouring control system consists of both electrical and mechanical elements. The electrical elements are the electronic amplifiers, networks, amplidyne, and motor, and the mechanical elements are the gears, shafts, lead screws, and machine slides which compose the mechanical drives of the machine tool. In determining the performance of a contouring control system, each of these elements must be considered by itself and in relation to the other elements of the system.

One-Dimension Contouring Control System

GENERAL THEORY OF OPERATION

The terminology "1-dimension" is applied to contouring controls which govern motion along only one machine axis. A typical 1-dimension contouring control system is shown in Fig. 1, and the block diagram of the system in Fig. 2. The input to the system, the command, is the template contour. This contour is sensed, or measured, by the tracer-head stylus. Thus the position of the stylus itself, or the position of an element actuated by the stylus, is the reference input of the system. The output of the system, the controlled variable, is the position of the cutter.

The cutter position is fed to the tracer head by mechanical feedback elements and serves as the primary feedback signal. The tracer head compares the position of the stylus with the position of the tracer head and the difference between these two is the position error. This error is the actuating signal of the system. In addition, the tracer head converts the physical position error into an electrical error signal. The error signal is subsequently amplified and modified by net-

works, the control elements. The amplified and modified electrical error signal is designated the manipulated variable and is the input signal of the speed loop.

For the sake of convenience, the input to the speed loop has here been designated the manipulated variable, since it is enclosed only in the primary feedback loop. Signals beyond the input to the speed loop might also be considered the manipulated variable, but these are enclosed in secondary feedback loops.

The speed loop consists of additional amplifiers, networks, and a motor-amplidyne combination, and employs both tachometer and motor armature voltage feedback. Over its linear range, the speed loop produces a motor speed directly proportional to the modified error signal. The motor is coupled to the machine slide by means of gears, shafts, and a lead screw and nut. The controlled system thus consists of the speed loop and the mechanical drive of the machine tool.

SPEED LOOP

The elements of the speed loop are identical for all types of General Electric contouring control systems, whether 1-dimension, 2-dimension, or other. The speed loop is shown in more detail in Fig. 3. Feedback is used to improve the dynamic response and linearity of the system, and to produce a motor speed directly proportional to the speed loop input. As the block diagram indicates, this is accomplished by employing a motor-speed feedback signal produced by a tachometer coupled to the feed motor. In some cases, the tachometer is directly connected to the motor shaft, while in others the tachometer is geared or belted to the motor shaft or some auxiliary shaft. As indicated, the tachometer voltage is filtered by a conventional 2-section resistance-capacitance filter before it is fed into the system.

In the steady state, the stability networks and the armature voltage feedback may be neglected since they function only transiently. Any difference which exists between the input to the speed loop and the speed feedback signal is a speed error signal. This signal is amplified and subsequently used to excite the control fields of an amplidyne generator. Over the linear range, the am-

plidyne armature voltage is directly proportional to the speed error signal and is applied to the armature of a d-c shunt motor with constant field excitation. The open loop gain of the speed loop is conventionally high, in the order of 1,000, and consequently the speed error is small, about 0.1 volt.

In practice, the speed loop is initially stabilized without the use of armature voltage feedback by employing only networks G_1 and G_2, which are shown in detail in Fig. 4. The amplifier, which has an adjustable gain K, is merely shown as a block. The armature feedback elements are shown by a dashed line since they are assumed disconnected for the present. The over-all transfer function for the networks and the amplifier is

$$\frac{E_o(p)}{E_i(p)} = \frac{K(T_1p+1)(T_2p+1)}{T_1T_2p^2+[T_1+T_2+T_{12}(1+K)]p+1} \quad (1)$$

where

$$T_1 = R_1C_1$$
$$T_2 = R_2C_2$$
$$T_{12} = R_1C_2$$

Typical values are

$T_1 = 0.055$ second
$T_2 = 0.085$ second
$T_{12} = 0.275$ second
$K = 51 = 34.2$ decibels (db)

The asymptotes of the amplitude portion of the transfer function are plotted in Fig. 5. As a result of the shape of the plotted transfer function, the name "notch network" has been applied to this circuit. The important aspects of the notch are the bottom corner frequencies, ω_1 and ω_2, and the depth d of the notch. The general expression for the notch depth is

$$d = \frac{R_2}{R_2+R_1(1+K)} \quad (2)$$

which is the expression for the transfer function of a simple voltage divider consisting of R_2 and $R_1(1+K)$ in series.

For the network of Fig. 5, the depth is 1/169, or −44.7 db.

The over-all gain of the network and the amplifier in the frequency range between ω_1 and ω_2 is

$$\text{Over-all gain} = \frac{KR_2}{R_2+R_1(1+K)} \quad (3)$$

If K is much greater than 1, and KR_1 is much greater than R_2, this expression reduces to

$$\text{Over-all gain} = \frac{R_2}{R_1} \quad (4)$$

Equation 4 indicates that in the fre-
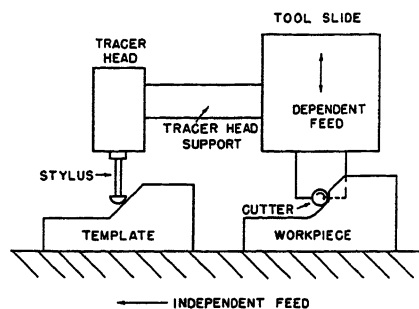
Fig. 1 (left). Sketch of 1-dimension contouring control system

Fig. 2 (right). Block diagram of 1-dimension contouring control system

network appreciably modify the speed loop locus in the vicinity of the point $-1+j0$. A decrease in the loop gain may actually cause the locus to move closer to the point $-1+j0$, thereby resulting in a less stable system. Three typical response curves which illustrate this condition are shown in Fig. 8. These curves were taken on a Brush recorder.

In some instances, the operation is sufficiently linear over the entire speed range and the response curves at various amplitudes are almost identical.

No attempt is made to select the parameters of the stability network prior to the installation of the control on the machine. In general, insufficient information is available about the characteristics of the machine prior to the installation. In the past, such things as irregular friction along the machine axis, tachometer belt stretch, and deflection or tilting of the machine elements have made stabilization of the speed loop difficult, and impossible to predict.

The speed of response is to some extent determined by the ratio of the amplidyne rating to that of the motor. During acceleration, large currents may be required by the motor. If these cannot be supplied by the amplidyne, the speed of response will be impaired. From a practical point of view, considering both economical factors and performance, the kilowatt rating of the amplidyne should be twice the horsepower rating of the motor. Where this ratio has been maintained and 3,600-rpm amplidynes have

quency range between ω_1 and ω_2 the gain remains constant regardless of how the amplifier gain K is varied. Therefore, the d-c gain of the speed loop may be changed without appreciably changing the gain of the loop near the crossover frequency. In most instances, it has been possible to change the gain by a factor of 10 without producing any noticeable change in the response of the speed loop.

In practice, when the speed loop is being stabilized, the amplifier gain K is set at its maximum value and the values of R_1, C_1, R_2 and C_2 are adjusted to obtain the best possible transient response to step functions applied to the input of the loop. The components are adjusted to obtain an overshoot of perhaps 20–30 per cent (%) on input step functions of various amplitudes between 10% and 100% of the maximum motor speed. A typical response to a 25%-speed step function is shown in Fig. 6. Note the large overshoot and the undershoot.

A sufficient amount of the armature-voltage feedback signal is then introduced into the system to reduce the initial overshoot to 5% to 10% and to completely eliminate the undershoot. The typical response of a properly stabilized speed loop is shown in Fig. 7. This response was measured on the same system as that of Fig. 6. Note that the overshoot is smaller and that the undershoot no longer exists. The time required to reach 63% of the steady-state speed is essentially the same in both cases.

Ideally, the response should be the same regardless of the amplitudes of the applied step functions. However, owing to the nonlinearities in the system, this is not always the case. In some instances, the overshoot is greater and the speed of response faster when small amplitude input step functions are applied. The slower response at higher signal levels is due to the decrease in loop gain which results when saturation occurs either in the amplidyne or the amplifiers.

Under some conditions, the overshoots may be greater at the higher speed steps. This usually occurs when the phase shifts introduced into the loop by the stability

been used, the response time of the speed loop has been approximately 0.05 second for motors ranging from 3/4 to 20 horsepower.

THE TRACER HEAD

The tracer head senses the contours of the template with the stylus, and also measures the position of the cutting tool, since in most cases the tracer head is mounted on the tool slide or on a support attached to the slide as shown in Fig. 1. When the position of the cutter is not in correspondence with the contour of the template, a mechanical error exists. The 1-dimension tracer head converts the mechanical error into a d-c electrical error signal. This is shown in Fig. 9. The gain of the tracer head is 2 volts d-c per 1/1000 inch of mechanical error. The tracer head output is in phase with the mechanical position error up to frequencies far beyond those to which the system can respond. Therefore, for all practical purposes, it can be considered a nonfrequency sensitive element.

If either the tracer head support, represented by the "feedback elements" of Fig. 9, or the template, is not rigid, vibration due to the spindle motors or due to the actual cutting may be transmitted to the tracer head. As a result, the vibration will appear in the electrical error signal at the tracer-head output. The system, in attempting to follow the vibra-
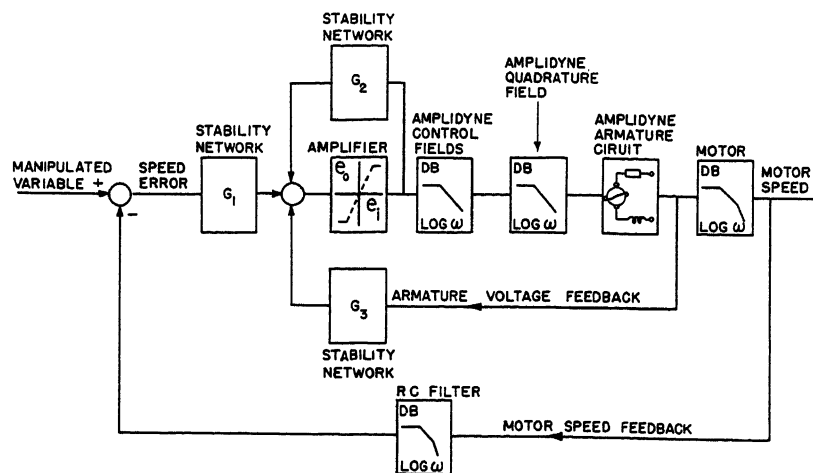
Fig. 3. Block diagram of speed loop

Ellert—Feedback in Contouring Control Systems

Fig. 4. Speed loop stability network

$\omega_1 = \dfrac{1}{T_1} = 18.2 \ \dfrac{RAD.}{SEC.}$

$\omega_2 = \dfrac{1}{T_2} = 11.8 \ \dfrac{RAD.}{SEC.}$

+34.2 DB

−10.5 DB
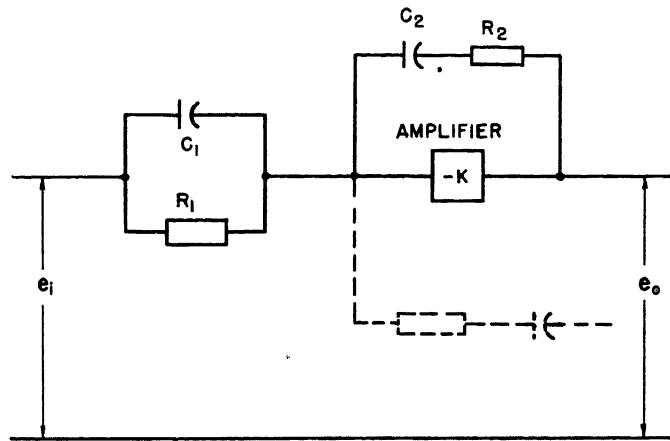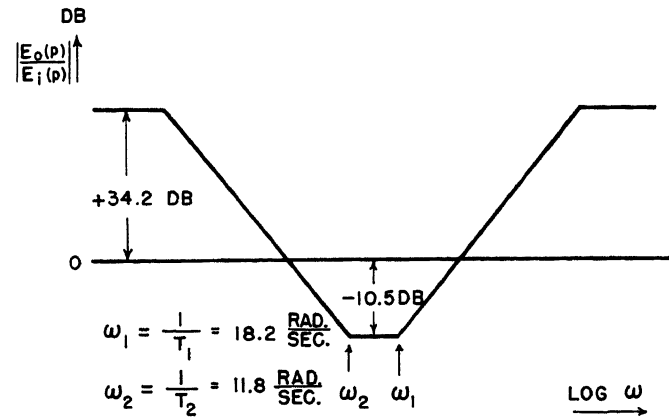
Fig. 5. Transfer function of the speed loop stability network

tion signals, may produce an unsatisfactory surface finish on the workpiece.

The surface finish can be improved considerably by the use of an amplitude sensitive filter network which will be discussed in detail later. However, the proper remedy for this condition is a redesign of those portions of the machine responsible for the vibration or its transmission to the tracer head. This solution does not degrade other aspects of the over-all performance in attempting to remedy this undesirable condition.

The 1-dimension tracer head also contains two limit switches. One of these, designated as the "automatic takeover limit switch," transfers the operation from manual to automatic when the stylus has made contact with the template and been deflected a prescribed amount. The other limit switch, designated the "overdeflection limit switch," stops automatic operation (contouring) when the stylus has been deflected beyond the normal operating region. This protects the tracer head, the machine, and the workpiece.

POSITION LOOP: THEORY OF OPERATION AND STABILIZATION

The basic position loop of the 1-dimension contouring control system consists of the tracer head, a group of amplifiers with adjustable gain, the speed loop, and the mechanical drive of the machine tool. As stated previously, a voltage is required at the input to the speed loop to produce a given motor speed. Because of the nature of the system, a voltage will be applied to the input of the speed loop only when an error exists between the position of the cutter and the contour of the template. This immediately indicates that the system will exhibit a zero position error with a constant position input and a constant position error with a constant velocity input.

When the cutter is in correspondence with the contour of the template, the position error and the tracer-head output are zero. Therefore, the signal applied to the input of the speed loop is also zero. If no unbalance exists in the system, the motor will be at a standstill. This condition exists when the system is following a zero-degree slope on the template (constant position input), that is, when the path followed on the template is exactly perpendicular to the axis of the controlled motion. However, if the stylus is instantaneously deflected beyond this null position, a position error will exist transiently. The resulting error signal will be applied to the speed loop causing the motor to rotate in such a direction as to return the stylus deflection to the null position and thereby reduce the error to zero. The rate at which the motor attempts to reduce the error is proportional to the magnitude of the error itself. Thus, the system attempts to maintain zero position error if possible.

If the stylus is following a constant slope on the template (constant velocity input), the tracer head must be moved in order to maintain a constant stylus deflection, and thereby produce a constant error signal. For example, if the system is to follow up a 45-degree slope, the tracer head (and the cutter) must be raised at the same rate that the tracer head is moved along the template by the independent feed of the machine. If the tracer head is not raised rapidly enough, the stylus deflection will continue to increase, resulting in an increasing error signal and a correspondingly increasing motor speed, or dependent feed speed. Equilibrium is attained when the error signal becomes constant at the value necessary to produce the required dependent feed rate. For any given slope on the template, the greater the independent feed rate along the template (and

workpiece), the greater the dependent feed rate required perpendicular to the template, and also the greater the position error.

The position error, in most cases, can be reduced to an acceptable value by increasing the open-loop gain of the position loop. However, it will be found that increasing the open-loop gain beyond a certain point will cause the system to become unstable. For the basic system, with no stability network in the position loop, the coefficient relating dependent feed speed and the position error may vary between 0.5 and 1.0 inch per minute per 1/1000 inch, or between 8 and 16.7 seconds^{-1}. This coefficient is called the "velocity error coefficient." It must be clearly understood that this coefficient relates position error and dependent feed speed (or motor speed). The terminology velocity error coefficient has led some to believe that a steady-state velocity error exists between the input and the output of the system and that this coefficient is a measure of the steady-state velocity

MOTOR SPEED

20% OVERSHOOT

T = 0.06 SEC.   25% SPEED

TIME
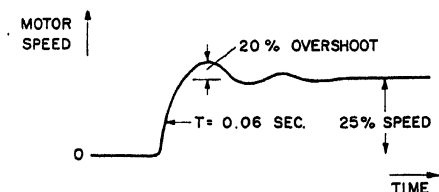
Fig. 6. Speed loop transient response without armature voltage feedback

MOTOR SPEED

10% OVERSHOOT

T = 0.06 SEC.   25% SPEED

TIME
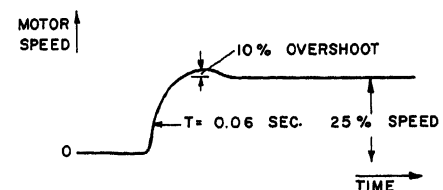
Fig. 7. Speed loop transient response with armature voltage feedback

Ellert—Feedback in Contouring Control Systems
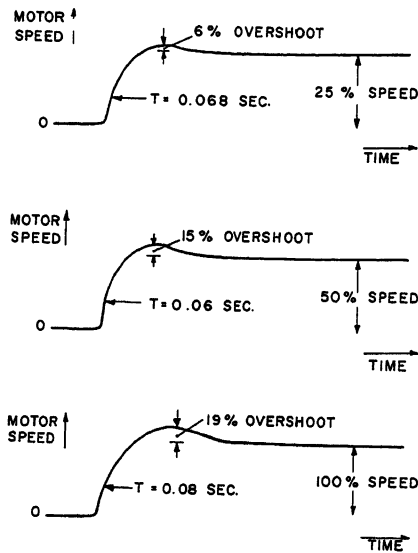
Fig. 8. Nonlinear speed loop response curves

error. It should be clear that this concept is not correct.

The position loop of the basic system is readily stabilized by increasing the gain of the amplifiers following the tracer head to the point where the desired position loop transient response is obtained. If the gain is increased beyond this point, the system becomes unstable. The resonant frequency of this system varies between 1 and 6 cycles per second, depending primarily on the amplidyne and motor time constants.

If the position error obtained with the basic system is too large for a particular application, it will be necessary to increase further the gain of the position loop. This can be accomplished without resulting in an unstable system if a stability network is added to the position loop. The network is shown in the block diagram of Fig. 9 and also in Fig. 10. This is a con-

ventional notch network and its transfer function is

$$\frac{E_0(p)}{E_i(p)} = \frac{(T_1p+1)(T_2p+1)}{T_1T_2p^2+(T_1+T_2+T_{12})p+1} \quad (5)$$

where

$T_1 = R_1C_1$
$T_2 = R_2C_2$
$T_{12} = R_1C_2$

Typical values for these time constants are

$T_1 = 0.05$ second
$T_2 = 0.1$ second
$T_{12} = 0.17$ second

The complete transfer function is plotted in Fig. 11. In most instances, the resonant frequency ω_r of the system is found to occur near the second lead break of the network ω_1. This region is cross-sectioned in Fig. 11. The lead portion of the network is thereby employed to offset the effects of one of the lag breaks of the system.

For this network, the notch depth d is expressed by the relationship

$$d = \frac{R_2}{R_1+R_2} \quad (6)$$

For a typical system, the notch depth was 1/5. This means that the addition of the network to the system allowed the position loop gain to be increased by a factor of approximately 5. The position error is decreased by this same factor. Theoretically, the gain of the system may be increased to almost any desired value merely by deepening the notch accordingly. However, there are other factors which may prohibit the use of high gains and the corresponding networks. These factors will be discussed later in this paper.
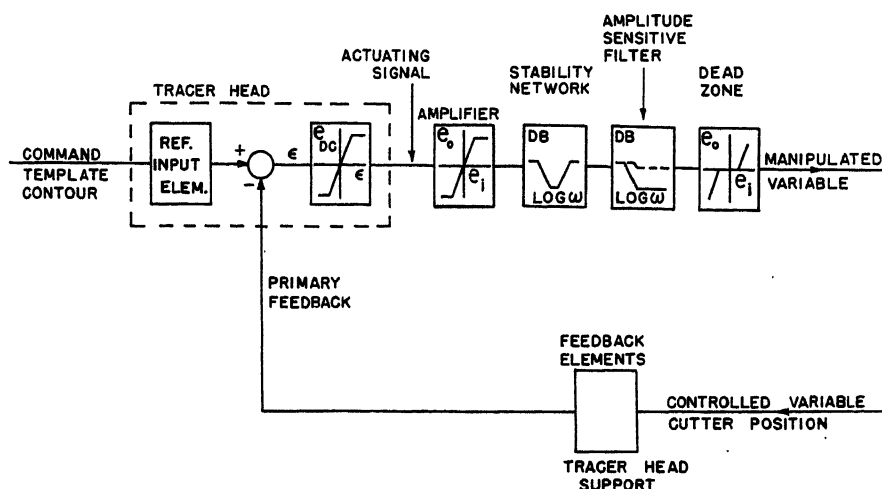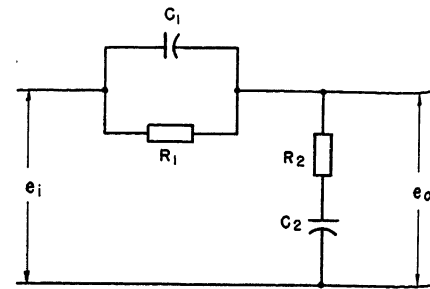
Fig. 10. Position loop stability network

Velocity error coefficients of 4 inches per minute per 1/1000 inch (or 67 seconds^{-1}) have been obtained with this system under normal conditions. In one case, where the mechanical characteristics of the machine were ideal, a velocity error coefficient of 17.2 inches per minute per 1/1000 inch (or 288 seconds^{-1}) was obtained.

As with the speed loop, the parameters of the position loop stability network are not selected until the control is installed on the machine because the variations in the characteristics of the machine are great and, in most cases, unknown. The actual stabilizing of the position loop by adjustment of the stability network parameters follows a logical procedure. Only a knowledge of the general shape of the open position loop transfer function and the transfer function of the network is required. This will not be discussed in this paper.

EFFECTS OF BACKLASH ON PERFORMANCE

Backlash exists to some degree in the mechanical drives of all machine tools. It can be reduced somewhat by the use of precision gears as well as adjustable anti-backlash nuts on the lead screws of the machine. However, as yet it has not been possible to eliminate the backlash completely.

In 1-dimension contouring control systems, the backlash is enclosed in the position loop as shown in Fig. 12. This arrangement is advantageous since it allows the system to provide some compensation for the backlash. In those systems where the backlash is external to the position loop, the controlled variable may differ from the command by the amount of the total backlash. If the backlash is excessive, the resulting position error may not be acceptable.

However, the existence of backlash inside the position loop presents some formidable problems, the most important of which is the effect of backlash on the stability of the position loop. Under conditions of high open-loop gain and corresponding deep-notch networks, the

Fig. 9. Block diagram of input portion of 1-dimension contouring control system position loop

Ellert—Feedback in Contouring Control Systems
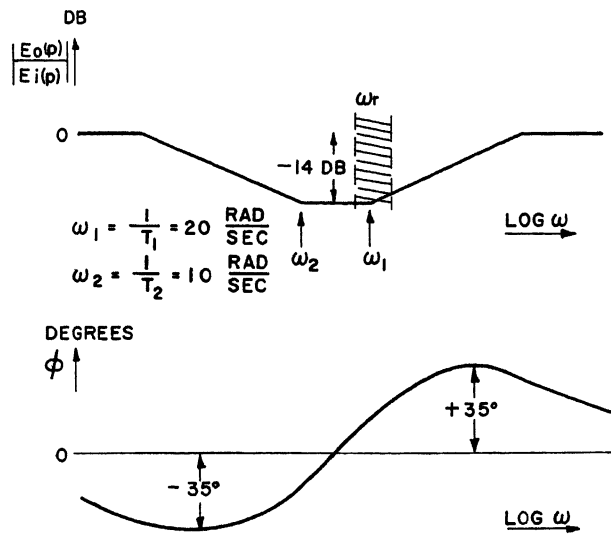
Fig. 11 (above). Transfer function of position loop stability network
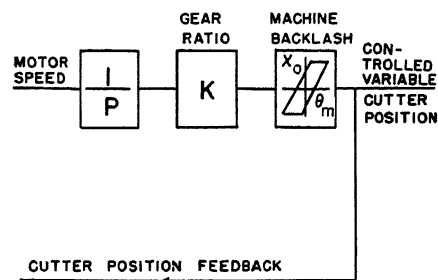
Fig. 12 (left). Block diagram of output portion of position loop

Fig. 13 (right). Backlash locus and typical system loci
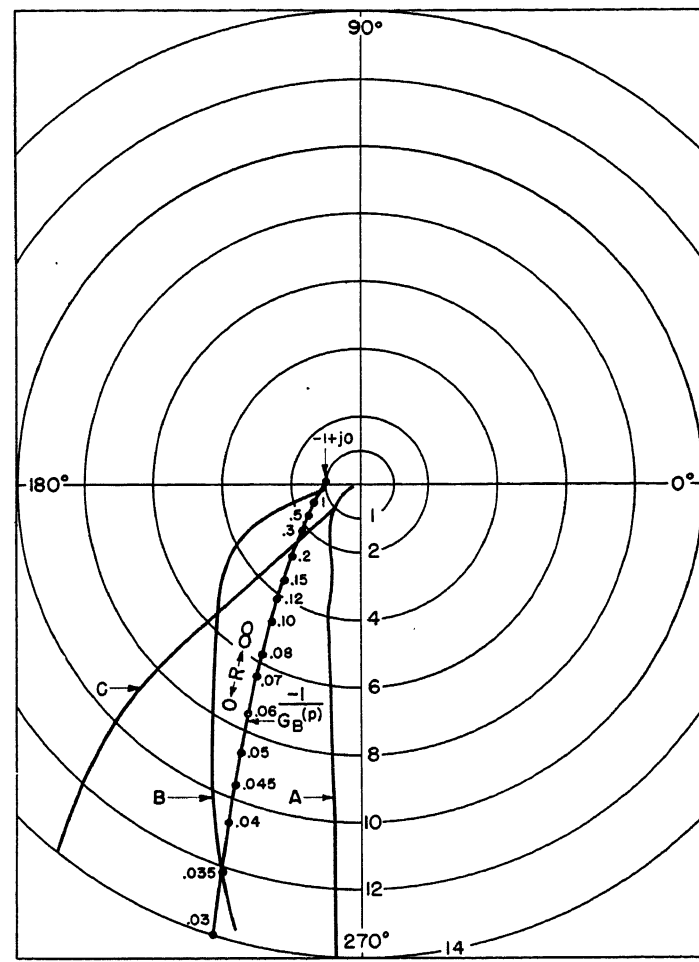
backlash may cause the position loop to become unstable when the system is attempting to hold zero error in the backlash region along a zero-degree slope. This phenomenon is best understood by considering the open-loop transfer function without backlash and the describing function of the backlash. To a first order of approximation, the describing function for the backlash is

$$G_B(p) = \frac{X_0(p)}{X_i(p)} = \frac{1}{\pi}\left(\pi - \beta + \frac{1}{2}\sin 2\beta - \frac{\sin^2\beta}{\omega}p\right) \quad (7)$$

where

$$\beta = \cos^{-1}\left(1 - \frac{X_B}{X_{mt}}\right) \quad (8)$$

X_{mt} is the maximum value of the displacement applied to the input of the backlash element, and X_B is the total backlash.

One conventional method of determining the stability of systems containing backlash is to plot the open-loop transfer function of the system without backlash, and the negative reciprocal of the backlash describing function $-1/G_B(p)$, on polar co-ordinate paper.

If the two loci intersect, the system

may exhibit a backlash hunt whose frequency is the frequency of the system transfer function at the point of intersection, and whose amplitude is determined by the backlash and the value of the backlash describing function at the point of intersection. The approximate locus for $-1/G_B(p)$ is plotted in Fig. 13 in terms of a dimensionless quantity R, where

$$R = \frac{X_{mt}}{X_B} - \frac{1}{2} \quad (9)$$

The values of R are indicated along the plotted locus. An open-loop transfer function which is representative of a basic 1-dimension contouring control system with no position loop stability network is also plotted in Fig. 13, curve A. The velocity error coefficient of this system is 15 seconds^{-1}.

Inspection of Fig. 13 indicates that the basic system should not exhibit a backlash hunt since no intersection between the two loci exists. Experimental data on a number of 1-dimension contouring control systems verifies this conclusion, since not one of them exhibited a backlash hunt under these conditions. However, if the gain of the system is increased by a factor 5, the open-loop transfer

function intersects the backlash locus, as shown in Fig. 13, curve B. This system would exhibit a backlash hunt if it were not already unstable as a result of its proximity to the point $-1+j0$. Inspection of Fig. 13 indicates that a reduction in gain by a factor of 3 or 4 would probably eliminate the backlash hunt. If a notch network with a depth of 1/5 is introduced into the position loop to stabilize the system at the higher gain level, the resulting open-loop transfer function is modified to that shown in Fig. 13, curve C. An intersection again exists between the backlash locus and the system locus, causing the system to exhibit a backlash hunt. The intersection occurs at $\omega = 8$ radians per second, which will be the frequency of the backlash hunt. The value of R at the intersection is 0.27. If we assume a total backlash of 0.010 inch, the amplitude of the backlash hunt will be 0.0027 inch, or $X_B R$.

Unfortunately, the position loop stability network has so modified the system locus that an appreciable gain reduction is required, to eliminate the intersection of the two loci. Under these conditions, the only alternatives would be to operate the system at a low value of open-loop gain with no position loop stability net-

$$\frac{E_0(p)}{E_i(p)}$$

Fig. 14. Transfer function of dead-zone circuit

As the foregoing indicates, a compromise must always be made between the required system response, the allowable position error, and the dead-zone width required to eliminate the backlash hunt. Unfortunately, the best compromise is a function of the characteristics of the machine and the particular requirements of the ultimate user of the contouring control system. As a result, this compromise can only be made in the machine tool manufacturer's plant or in the ultimate customer's plant. Therefore, no attempt is made to set the width of the dead zone prior to the installation of the control on the machine.

SURFACE FINISH

As mentioned previously, when excessive machine vibration is transmitted to the tracer head, the vibration will appear in the electrical error signal at the tracer-head output. If the vibration components of the error signal are not filtered out, the surface finish of the workpiece may not be acceptable. However, if a conventional filter were added to the system, the system response would be impaired to the point where it would be unsatisfactory.

The 1-dimension contouring control system makes use of what might be called an "amplitude-sensitive filter," That is, when signals of small amplitude are applied to the filter, it performs as a low-pass filter. However, for large amplitude signals, it ceases to attenuate the higher frequency signal components. This circuit is shown in Fig. 9 where the solid characteristic exists for error signals of small amplitude, and the dashed one for signals of large amplitude.

The filter circuit and the amplitude portion of its transfer function are shown in Fig. 15. The expression for the transfer function is

$$G_F(p) = \frac{E_0(p)}{E_i(p)} = \frac{(RCp+)1}{(R+R_T)Cp+1} \quad (12)$$

For applied signals of small amplitude, the resistance of the Thyrite R_T is large and the attenuation is high, whereas for large amplitude signals, the Thyrite's resistance is extremely small. Therefore, for normal slopes, where the error signal is small, the signal is adequately filtered. During transients at corners where the error signal is large, the filter has essentially unity gain at all frequencies and the required system response is maintained.

Unfortunately, the filter affects the backlash hunt in much the same manner as the position loop stability network, i.e., the filter also introduces a phase lag

work, or to set up the workpiece so that no cuts are made along a zero-degree slope.

However, the addition to the system of a dead-zone or threshold circuit of adjustable width has in many cases made possible the elimination of the backlash hunt. The dead zone is shown in Fig. 9. A describing function may also be obtained for the dead zone by making a Fourier analysis of the output voltage of the dead zone when a sinusoidal voltage of amplitude E_m and frequency ω is applied to the input. If the ratio is taken of the fundamental component of the output voltage to the input voltage, the approximate describing function is obtained.

The expression is

$$G_{dz} = \frac{E_0(p)}{E_i(p)} = \frac{1}{\pi}(\pi - 2\alpha + \sin 2\alpha -$$

$$4 \sin \alpha \cos \alpha) \quad (10)$$

where

$$\alpha = \sin^{-1} \frac{E_{dz}}{E_m} \quad (11)$$

E_{dz} is one-half the width of the dead zone in volts. This expression indicates that there is no phase shift introduced into the system by the dead zone. The result is merely a loss of gain which is a function of the amplitude E_m of the signal applied to the dead zone and the dead-zone width. The curve which relates the dead zone gain to the ratio E_{dz}/E_m is plotted in Fig. 14. As the curve indicates, for small error signals, the open-loop gain is low, and the system locus will not intersect the backlash locus; therefore, no backlash hunt exists. During transients, as at corners on the template, the error signal rapidly rises to values which result in a dead-zone gain of essentially unity. Thus, the response of the system is not appreciably impaired.

On large slopes or when high independent feed speeds are used, the error signal is again large enough to make the gain of the dead zone essentially unity and thereby not produce any appreciable increase in the error. In a typical application, a ±2-volt dead zone was employed in a system where the total backlash was 0.011 inch. This dead-zone voltage, when referred back to the actual position error, was equivalent to ±0.0003 inch of error. This additional error was not an objectionable error even at the lower feed rates, where the error determined by the velocity error coefficient was itself low. At the higher feed rates, the error resulting from the existence of the dead zone in the system is only a small fraction of the total error.

The addition to the position loop of a stability network which introduces an appreciable amount of phase lag into the system may make the elimination of the backlash hunt through the use of the dead zone extremely difficult. As is indicated by the shape of curve C of Fig. 13, the decrease in gain required of the dead zone to eliminate the intersection of the loci may be large. This means a wide dead zone is required which, to some degree, will impair the system response and increase the position error. If an extremely deep notch network is used which introduces enough phase lag into the system to bend the locus further up toward the -180-degree region, a condition may be reached where the decrease in loop gain produced by the dead zone will actually increase the amplitude and frequency of the backlash hunt by causing the intersection of the two loci to occur higher up on the backlash locus. This has been found true in a few applications where extremely high velocity error coefficients were required.

Fig. 15 (left). Circuit and transfer function of amplitude sensitive filter

Fig. 17 (above). Block diagram of mid-portion of 2-dimension contouring control system

$$\omega_2 = \frac{1}{T_2} = 2 \quad \frac{RAD.}{SEC.}$$

$$\omega_T = \frac{1}{T_T} = 0.875 \quad \frac{RAD.}{SEC.}$$

$$\omega_T^1 = 1.98 \quad \frac{RAD.}{SEC.}$$

into the system which may require the use of a wider dead zone to eliminate the backlash hunt. Therefore, this factor must also be considered when attempting to arrive at the best compromise for overall system performance.

System Performance

As the foregoing discussions indicate, a statement of the performance of a 1-dimension contouring control is not a simple matter, since the performance is determined by a number of interrelated and indeterminate factors. However, a few general statements can be made.

The steady-state error is determined by the velocity error coefficient of the system. Four inches per minute per 1/1000 inch is a representative value for this coefficient. The position error ϵ when following a constant slope may be determined from the relationship

$$\epsilon = \frac{S_i \tan \beta}{K_p} \tag{13}$$

where S_i is the independent feed speed, β is the angle of slope, and K_p is the velocity error coefficient. For example, if the system is following a 45-degree slope at an independent feed speed of 20 inches per minute, the position error will be 0.005

inch if the velocity error coefficient is 4 inches per minute per 1/1000.

The transient error of the system at a corner on the template is determined by the effective radius of the corner, the speed of response of the position loop, the independent feed speed, the slope of the template following the corner, and the machine backlash in those cases where a reversal of direction is involved. The speed of response is in turn determined by the characteristics of the motor and amplidyne, the characteristics of the stability network used, and the amount of vibration filtering employed. Any data presented here would be of little significance unless accompanied by detailed information on the characteristics of the machine elements which, unfortunately, is not readily available.

Experience has shown that if the machine backlash is much over 0.010 inch, it is extremely difficult to eliminate the backlash hunt through the use of the dead zone without impairing the response of the system and increasing the steady-state error. Values of backlash below 0.005 inch are desirable.

In most cases, where the vibration has not been excessive (and this term is difficult to define), the surface waviness on aluminum has been below 60 microinches.

One-Dimension Contouring Control System Using Selsyns

In some applications, especially on large machines where it is not practical to connect mechanically the tracer head to the tool slide, a different system is used to obtain the electrical error signal. In such a case, one selsyn, or synchro, is driven from the stylus by means of a rack and

Fig. 16. Block diagram of input portion of 2-dimension contouring control system
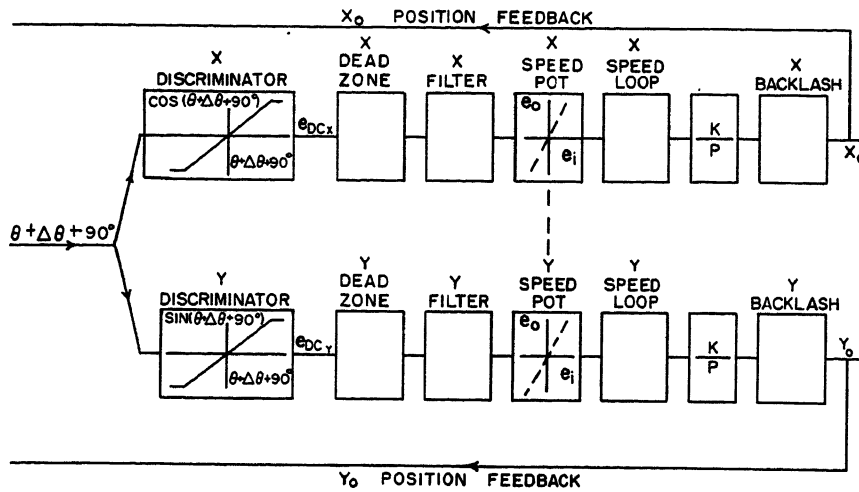
Fig. 18. Block diagram of output portion of 2-dimension contouring control system

pinion and perhaps an associated gear train. Another selsyn is driven from the tool slide by a similar rack-and-pinion arrangement, or is driven through gearing directly from the drive motor. In the latter case, the backlash between the motor shaft and the tool slide is not enclosed in the position loop and cannot be compensated for by the system. The rotor of one of the selsyns is excited by a 400-cycle-per-second signal from an oscillator, and the 3-wire stators of the two selsyns are tied together. The amplitude of the a-c voltage which appears at rotor of the other selsyn is a measure of the angular position error between the two selsyn rotors. This a-c signal is applied to an amplitude discriminator which produces a d-c voltage directly proportional to the sine of the angular error between the two selsyn rotors. The discriminator output corresponds to the error signal of the conventional system. The speed loops of the two systems are identical.

All that has been said about the conventional system concerning stabilization, the effects of backlash, surface finish, and performance applies equally well to this modified system. One additional factor must be considered, and that is the importance of the proper design of the selsyn-driving mechanism. If precision racks, pinions, and gears are not used, the selsyns will not run smoothly and excessive vibration signals at the discriminator output will result. Also, additional backlash may be introduced into

the system. These undesirable factors can be reduced by careful design of the mechanical elements involved.

Two-Dimension Contouring Control System

THEORY OF OPERATION

A 2-dimension contouring control system, as its name implies, simultaneously controls motion along two mutually perpendicular machine axes. A block diagram of the system, subdivided into three parts, is shown in Figs. 16 through 18. The input to the system is again the contour of a template, in this case a 2-dimensional template, as shown in Fig. 19. The template is mounted on the same base as the workpiece and, for this discussion, it is assumed that this base is stationary. The tracer head is held by a support attached to the tool slide, which is positioned simultaneously along the two mutually perpendicular machine axes by the feed motors.

If it is assumed that the friction between the stylus tip and the template is negligible, as is essentially the case when a ball-bearing tip is used, the only force exerted on the stylus is normal to the template. This is illustrated in Fig. 19, where the surface of the template is at an angle β with respect to the x-axis of the machine and the stylus deflection X_d is at an angle β with respect to the y-axis. Since the stylus tip is in contact with the template (the input), and the tracer head is attached to the tool slide (the con-

trolled variable), the stylus deflection is the position error of the system. This is shown in the block diagram of Fig. 16.

The tracer head has two mutually perpendicular electrical axes which must be aligned with the machine axes. Under this condition, the system error, which consists of a magnitude X_d and an angle β, is resolved into components parallel to the X and Y axes. The tracer head converts these deflection components into two inductances

$$L_x \propto X_d \sin \beta \qquad (14)$$

$$L_y \propto X_d \cos \beta \qquad (15)$$

Each of these inductances is connected in a separate bridge circuit. One bridge is excited by the 60-cycle-per-second reference voltage and the other is excited by the 90 degree leading quadrature voltage. Each bridge circuit is adjusted so that zero stylus deflection parallel to the corresponding electrical axis of the tracer head results in an inductance which exactly balances the bridge. The bridge output is then zero. If the stylus is deflected at some angle to the two tracer head axes, the output voltages of the two bridges, e_{ACX} and e_{ACY} are essentially

$$e_{ACX} = (K_x X_d \sin \beta) \sin \omega t \qquad (16)$$

$$e_{ACY} = (K_y X_d \cos \beta) \sin (\omega t + 90 \text{ degrees}) \qquad (17)$$

Actually each of these voltages may be phase-shifted an additional amount owing to the bridge circuit, but this is compensated for in the system and will not be considered here. The two bridge output voltages are added vectorially to obtain a single a-c voltage which is

$$e_{AC} = K X_d \sin (\omega t + \theta) \qquad (18)$$

where

$$\theta = 90 \text{ degrees} - \beta \qquad (19)$$

As this expression indicates, the amplitude of this a-c voltage is directly proportional to the stylus deflection X_d, and the electrical phase θ is a function of β, the angle of the stylus deflection. Therefore, this a-c signal voltage contains both the amplitude and phase of the position error.

The a-c signal voltage is fed to two different portions of the system. Consider for the present that one portion of the system is disconnected at point A of Fig. 17, and therefore does not function. The a-c signal voltage is fed through a clipping or squaring amplifier, which transforms the sinusoidal signal into a square wave and thereby removes any error information contained in the am-
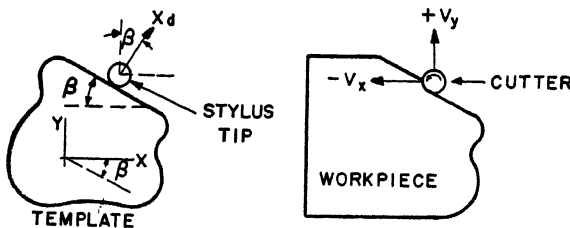
Fig. 19. Partial sketch of 2-dimension contouring control system

plitude of the signal. That is, the output voltage of the amplifier contains merely the phase β of the position error. This signal is then phase shifted 90 degrees by a phase shift circuit and applied to two phase discriminators. The output of the X-discriminator e_{DCX} is proportional to the cosine of the phase angle $\theta +90$ degrees of the applied signal and the output of the Y-discriminator e_{DCY} is proportional to the sine of that angle

$$e_{DCX} = K_1 \cos (\theta + 90 \text{ degrees}) = -K \sin \theta \quad (20)$$

$$e_{DCY} = K_1 \sin (\theta + 90 \text{ degrees}) = K_1 \cos \theta \quad (21)$$

These direct voltages are then fed through dead zones and surface-finish filters identical with those in the 1-dimension contouring control system. The X and Y speed potentiometers which are ganged together, as indicated by the dashed line connecting them, merely reduce these d-c signals to the proper value before they are applied to the X and Y speed loops.

The X and Y loops are identical in every respect to the speed loop in the 1-dimension contouring control system. The feed speeds, V_X and V_Y, resulting in the case of this 2-dimension contouring control, are

$$V_X = -K_2 \sin \theta = -K_2 \sin (90 \text{ degrees} - \beta)$$
$$= -K_2 \cos \beta \quad (22)$$

$$V_Y = K_2 \cos \theta = K_2 \cos (90 \text{ degrees} - \beta)$$
$$= K_2 \sin \beta \quad (23)$$

Referring these two velocities to Fig. 19 indicates that the stylus and cutter are moving up the slope. The resultant velocity of the stylus is the vector sum of the two component velocities

$$\bar{V}_R = \bar{V}_X + \bar{V}_Y = |V_R| / \underline{\alpha} \quad (24)$$

where α is the angle of the resultant velocity with respect to the X-axis. Substituting in the values for V_X and V_Y gives

$$|V_R| = K_2 \quad (25)$$

$$\alpha = \beta \quad (26)$$

Therefore the resultant velocity is tangential to the surface of the template and its magnitude does not depend on the slope of the template, i.e., it is not a function of β. The tangential velocity is determined by K_2, which can be set to any desired value by adjusting the speed potentiometers. It is also important to note that with a portion of the system disconnected at point A, the system can sense only the angle of the deflection. The amplitude X_d of the deflection has been removed from the system error signal by the squaring amplifier. There-

fore the system could follow a constant slope at a constant deflection X_d or it could just as well drift in or out, thereby varying the deflection. Ideally, it exhibits zero error to a constant velocity input.

Any deviation of the stylus deflection X_d from the desired value can be compensated for to some degree by measuring the amplitude of the a-c signal which is the combined output of the bridges. This is accomplished by reconnecting the circuit at point A. The rectifier and filter produce a direct voltage directly proportional to the amplitude of the a-c signal applied, thereby making e_{d-o} proportional to X_d. The direct voltage is applied through the conventional notch network to the bender circuit which produces a change $\Delta\theta$ in the phase of the signal passing through the phase shifter. Under these conditions, the resultant feed speeds are

$$V_X = -K_2 \cos (\beta + \Delta\theta) = -K_2 \cos (\beta + K_3 X_d) \quad (27)$$

$$V_Y = K_2 \sin (\beta + \Delta\theta) = K_3 \sin (\beta + K_3 X_d) \quad (28)$$

Therefore, if for some reason the deflection X_d of the stylus attempts to change from the preset value (usually 0.010 inch or 0.020 inch), the bender action will shift the phase of the error signal in such a direction as to oppose the change. The desired deflection is set by adjusting the bender circuit so that the change in phase $\Delta\theta$ is zero at the value of e_{d-o}, which corresponds to the desired deflection. The system will then attempt to maintain this constant deflection while following the template.

Unbalance or drift in the system will be partially compensated for by the bender action. However, under these conditions, a steady-state deflection error may exist. For a given change in phase required to compensate for an unbalance, the actual deflection error will be inversely proportional to the gain of the bender circuit K_B. The bender circuit is a part of the position loop and cannot have its gain increased indiscriminately, because the position loop may become unstable. As in the 1-dimension system, the stability network of Fig. 10 is used in the position loop to insure stability at high-loop gains. The stability network will not be discussed further here.

One additional factor concerning the deflection error must be considered. As lower feed rates are used (by setting the speed potentiometers to lower values), the position loop gain decreases and the position loop response becomes slower. This is not a desirable condition and has

been remedied by mechanically connecting the bender gain potentiometer and the speed potentiometers in such a manner that lowering the feed rate will increase the bender gain, thereby keeping the loop gain constant. This has proved very satisfactory.

System Performance

The remarks made previously concerning the effects of backlash and excessive machine vibration on the performance of the 1-dimension system apply equally well to the 2-dimension system. The steady-state errors, with constant velocity inputs, on the other hand, are in general smaller in the 2-dimension system, being in the order of 0.002 inch at the higher feed rates, and 0.0002 inch at the lower feed rates. The transient errors at corners are determined by the same factors as discussed in conjunction with the 1-dimension system.

Combinations of Basic Systems

Selective 2-Dimension Contouring Control System

A selective 2-dimension contouring control system is essentially a 2-dimension system in which motion along any two mutually perpendicular machine axes can be controlled simultaneously. The electrical circuits can be switched to control any two of the feed motors. One type of tracer head for this system contains three electrical axes and is sensitive to deflections along all three mutually perpendicular axes, but the signals are selected from only two at a time. In a new tracer head for this system, only two electrical axes exist and the stylus deflections along any two machine axes are mechanically switched to the same two electrical axes in the tracer head. The performance of this system is comparable to that of the conventional 2-dimension control.

Two-and-1-Dimension Contouring Control System

A 2-and-1-dimension contouring control system is used to control motion along three machine axes simultaneously. A 2-dimension system is used for profiling and a 1-dimension system for depth control. Two separate templates and tracer heads are used. The 2-dimension system supplies the independent feed for the 1-dimension system.

The performance on profiling is determined by the 2-dimension system, whereas the performance on depth control is determined by the 1-dimension

system. There is no electrical interaction between the systems except through auxiliary functions such as over- and under-deflection protection.

References

1. PRINCIPLES OF SERVOMECHANISMS (book), G. S. Brown, D. P. Campbell. John Wiley & Sons, Inc., New York, N. Y., 1948.

2. SERVOMECHANISMS AND REGULATING SYSTEMS DESIGN, H. Chestnut, R. W. Mayer. John Wiley & Sons, Inc., New York, N. Y., 1951.

3. THE EFFECTS OF BACKLASH AND OF SPEED-DEPENDENT FRICTION ON THE STABILITY OF CLOSED CYCLE CONTROL SYSTEMS, A. Tustin. *Journal*, Institution of Electrical Engineers, London, England, pt. IIA, 1947, pp. 143–51.

4. A FREQUENCY RESPONSE METHOD FOR ANALYZING AND SYNTHESIZING CONTACTOR SERVO-MECHANISMS, R. J. Kochenburger. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 270–84.

5. SINUSOIDAL ANALYSIS OF FEEDBACK-CONTROL SYSTEMS CONTAINING NONLINEAR ELEMENTS, E. Calvin Johnson. *Ibid.*, vol. 71, pt. II, 1952, pp. 169–81.

6. APPROXIMATE FREQUENCY-RESPONSE METHODS FOR REPRESENTING SATURATION AND DEAD BAND, H. Chestnut. *Transactions*, American Society of Mechanical Engineers, New York, N. Y., vol. 76, Nov. 1954, pp. 1345–63.

7. STABILITY CHARACTERISTICS OF CLOSED-LOOP SYSTEMS WITH DEAD BAND, C. H. Thomas. *Ibid.*, pp. 1365–82.

Quasi-Linearization Techniques for Transient Study of Nonlinear Feedback Control Systems

KAN CHEN
ASSOCIATE MEMBER AIEE

Synopsis: A new method of studying the transient response of a large class of nonlinear feedback control systems to a step input has been developed. The basis of the new method is quasi-linearization by which the nonlinear system is converted to an equivalent linear model, subject to certain limitations. Evaluation of the transient characteristics of the quasi-linearized system through linear servo techniques furnishes an insight to the problem of designing the nonlinear feedback control system to meet its prescribed transient response.

ANALYSES and syntheses of feedback control systems containing essential nonlinearities, whether inherent or deliberately put in the systems, have become a subject of wide interest. However, because of the lack of a general solution for nonlinear differential equations, all methods of nonlinear analysis are valid

only for certain types of nonlinear problems. When the input to a specific nonlinear system is given and the system transient response is to be found, one of the following methods may be used to obtain the result:

1. Performance of an actual test on the system with the given input.

2. Employment of an analogue or digital computer to represent the system and perform the test on the computer.

3. Use of a step-by-step procedure to calculate the approximate transient response.

However, if the problem is to design a nonlinear system which must have a prescribed transient response to a given input, these methods can be applied only in a strictly trial-and-error manner. The phase-plane method gives the designer an insight to the synthesis problem by presenting a perspective of all transient possibilities. Unfortunately, the consideration of adding a compensation network is usually rejected in this approach, because the phase-plane method cannot be simply extended to systems possessing many energy-storage elements.

A recently developed technique known as the describing function method[1-4] has proved very useful for designing as well as analyzing nonlinear feedback control systems for stability considerations. This method, based on quasi-linearization, converts the nonlinear element to an equivalent linear model whose parameters depend on the input magnitude and frequency. The method is particularly valuable because of its great

similarity to the well-known frequency-response method for designing linear servo systems. However, attempts to extend the describing function method to study the nonlinear transient problem have not appeared fruitful. Correlation between the frequency response and the transient response of a nonlinear system is unlikely in view of the invalidity of the principle of superposition in nonlinear analysis.

The method to be introduced by this paper is a counterpart of the describing function method as it is based on the same general philosophy of quasi-linearization but is developed for studying the transient rather than the stability of nonlinear feedback control systems. It differs from the describing function method in that a transient type of test function instead of a sinusoidal test function is used to obtain the quasi-linear model. The concept of simplicity is to be emphasized in the development of the method in order that the resultant scheme will particularly facilitate the design of nonlinear systems. Analysis of nonlinear systems is thus only a secondary function of the method.

Since there are various standards by which the transient performance of a control system may be judged, it is necessary at the outset to state the nature of the transient problem to be considered.

1. The input to the nonlinear system is assumed to be a step. This assumption, however, need not exclude the consideration of other forms of inputs which may be regarded as the step response of a fictitious network.

2. As the transient specifications of control systems are often in terms of the delay time, the rise time, the maximum overshoot and undershoot, the settling time, and the transient envelope of the system output, the description of system transients in this paper will be in terms of these transient characteristics.

3. Because of the approximation procedure involved in quasi-linearization, as will be described later, the type of nonlinearity to be considered is that of 2-terminal blocks containing piecewise linear gain followed by linear low-pass filter networks.

Fig. 1. Typical form of nonlinear feedback control system to be studied

Limited as the scope of the problem appears to be, a large number of practical systems fall within this scope with justified approximations. As the number of energy storage elements which exist in the linear portion of the system is immaterial, the range of application of the method is actually quite wide.

General Principle of Quasi-Linearization

Consider the nonlinear system in Fig. 1 in which the block N represents the nonlinear portion of the otherwise linear system. If the input θ_i is given, the response of the system at its various parts, $x_i(t)$, $x_0(t)$, and $\theta_0(t)$, are uniquely defined. Thus the nonlinear element N may be replaced by an equivalent linear network, provided the network bears an input-output relationship defined by $x_i(t)$ and $x_0(t)$. Obviously, the parameters of the equivalent network will in general be functions of $x_i(t)$. Hence the following principle of quasi-linearization may be deduced:

The quasi-linear representation of a nonlinear element should be obtained by testing the element with its actual input or with a time function which approximates the actual input as closely as possible.

Although not appearing previously in the literature, this principle has been followed by all present quasi-linearization
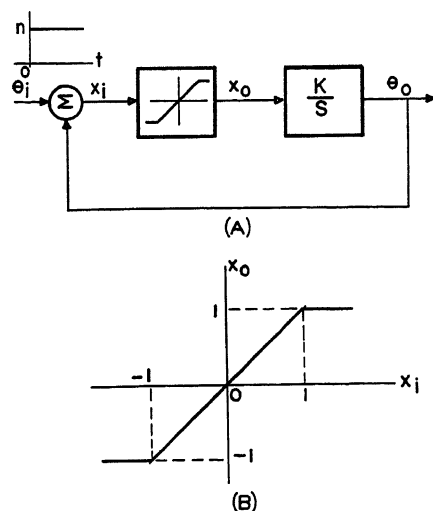
Fig. 2. Simple nonlinear feedback control system containing normalized limiter

techniques. For example, when the input to a nonlinear characteristic is known to be a small increment around a quiescent operating point, the nonlinear characteristic will be tested around its operating point to obtain its equivalent linear gain. This is the well-known linearization technique, which is a special case of quasi-linearization. When a nonlinear system is in the steady state, the input to the nonlinear element is approximately sinusoidal (magnitude may range from zero to infinity), if it is preceded by some low-pass linear filters. The equivalent linear transfer function of the nonlinear element under steady state may therefore be obtained by testing the nonlinear element with a sinusoidal input. This is the describing function method. Since the input to the nonlinear element is not sinusoidal in the transient state, the describing function does not represent the nonlinear element under transient. When the input to the nonlinear element has random magnitude which can be defined only by its statistical property, the test function for the nonlinear element should also be random and have the same statistical distribution in magnitude. This is the approach used by Booton in his quasi-linearization method.[5]

Quasi-Linear Model for Transient Study

In the light of the expressed principle the quasi-linear representation of a nonlinear element for transient study should be obtained by testing the nonlinearity with a transient type of signal which simulates the actual input to the nonlinear element. However, before the system transient is determined, the transient mode of the actual input to the nonlinear element is unknown. Hence the following step-by-step procedure of quasi-linearization is proposed.

Since the nonlinearity considered is piecewise linear, the operating point of the nonlinear element is expected to shift during the transient over several linear ranges. A quick evaluation of the actual transient response in the first linear range through linear servo techniques determines the test function for the nonlinearity in the first two linear ranges. The quasi-linear representation of the nonlinear system thus obtained permits a quick evaluation of the actual transient response in the second linear range. This will in turn determine the test function for the nonlinearity in the first three linear ranges, and so forth. Although this procedure appears to be similar to the ordinary step-by-step solution, it provides
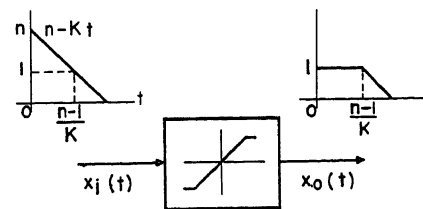
Fig. 3. Testing limiter by actual input

useful guidance for design work by indicating how system transient characteristics would vary with the system parameters, as will be shown in later examples.

Consider the simple nonlinear feedback control system in Fig. 2(A). The nonlinear portion of the system contains a normalized limiter whose characteristic is described in Fig. 2(B). The input $\theta_i(t)$ is a step function of magnitude n, which is greater than unity. Thus, immediately following $t=0$, the limiter output is unity and $\theta_0(t)=Kt$ until $t=(n-1)/K$. Hence, for $0<t<(n-1)/K$, the actual input to the nonlinear element is $x_i(t)=n-Kt$, and this function is used to test the limiter. The corresponding output is shown in Fig. 3. The nonlinear element can then be replaced by an equivalent active linear element as shown in Fig. 4(A). (The use of an equivalent passive linear element is possible but has proved not so flexible as the use of an active element.) The passive part of the quasi-linear network is the linearized gain of the limiter in the unlimited region ($=1$) and the active signal $a(t)$ accounts for the discrepancy between $x_i(t)$ and $x_0(t)$, the shaded portion of x_i in Fig. 4, when the limiter operates in the limited region. To facilitate analysis of the quasi-linear-
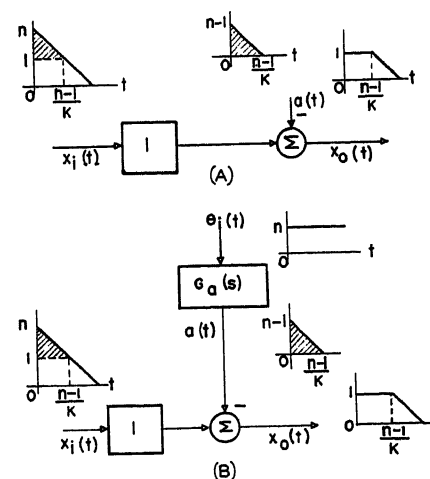
Fig. 4. Replacement of nonlinear element by quasi-linear element

A—Active quasi-linear element
B—Equivalent block diagram of A
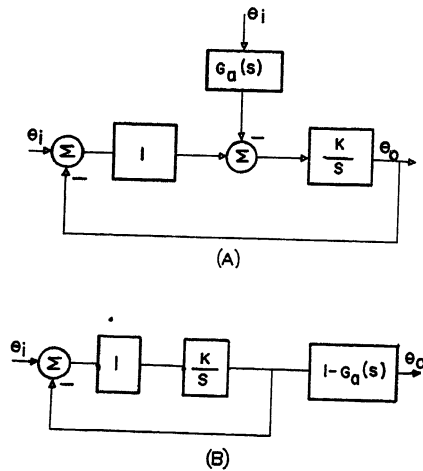
Fig. 5. Resultant quasi-linear representation of Fig. 2
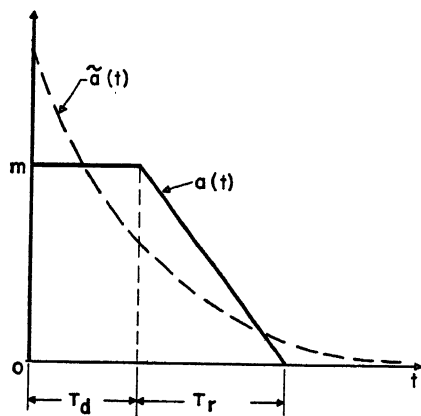
Fig. 6. Approximation of trapezoidal pulse by single exponential function

ized block diagram, Fig. 4(A) is transformed to Fig. 4(B), where the transfer function $G_a(s)$ is defined such that when its input is a step function of magnitude n its output is equal to $a(t)$. The mathematical expression of $G_a(s)$ is

$$G_a(s) = \frac{\dfrac{n-1}{s} - \dfrac{K}{s^2}\left(1 - \epsilon^{-\frac{n-1}{K}s}\right)}{n/s}$$

$$= \frac{n-1}{n} - \frac{K}{ns}\left(1 - \epsilon^{-\frac{n-i}{K}s}\right) \quad (1)$$

Replacement of the nonlinear element by the blocks in Fig. 4(B) results in the quasi-linearized system shown in Fig. 5(A), which may be converted to the form given in Fig. 5(B) by block diagram manipulation. The over-all transfer function is then

$$\frac{\theta_0}{\theta_i}(s) = \frac{K}{s+K}[1 - G_a(s)] \quad (2)$$

The effect of nonlinearity is thus approximated by inserting a certain linear network outside the closed loop. Notice that the parameters of $G_a(s)$ are functions of n as well as K, indicating that the shape of

the system response is dependent on the input magnitude. This is indeed a property of nonlinear systems.

The input to the limiter, $x_i(t)$, in the foregoing example was obtained by inspection. It is conceivable that its evaluation would be much more difficult if the linear portion of the system were more complex. $G_a(s)$ given in equation 1 is a transcendental function of s and hence requires special techniques in analyzing the resultant over-all system transfer function. For design and for synthesis purposes, it is always desirable to evaluate the input to the nonlinear element quickly even in a complex system and to approximate $G_a(s)$ by an algebraic function of s even at a sacrifice of some accuracy in the final result. This point of view leads to the two approximation schemes to be described.

Quick Evaluation of Approximate Transient Response of Linear Systems

The purpose of developing methods for quick evaluation of approximate transient response of linear systems in connection with the proposed quasi-linearization technique is twofold: To determine the proper test function for the nonlinear element, thereby obtaining quasi-linearization, and to appraise the equivalent linear system after quasi-linearization is performed. The methods to be developed contain no basically new ideas. They are merely a combination, a modification, and an extension of a number of known linear servo techniques.

Both the frequency response asymptote plot and the pole-zero plot of linear transfer functions are used as working media for evaluating the transient response of the system and for redesigning the system to meet prescribed transient response. Brief derivation of these methods is given in the Appendix. Only those results which are instrumental in working out the sample problems later in this paper are given in the Appendix. For other related results the reader is referred to the original work[6] from which the material is excerpted.

Pade's Approximation

It has been explained that the proposed quasi-linearization procedure requires the finding of an algebraic transfer function which approximately describes an arbitrary input-output relationship. A method known as the Pade's approximation can fulfill this function to any desired degree of accuracy.[7] It consists of
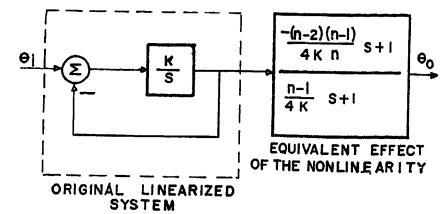
Fig. 7. Equivalent effect of nonlinearity as interpreted by quasi-linearization with proper approximation
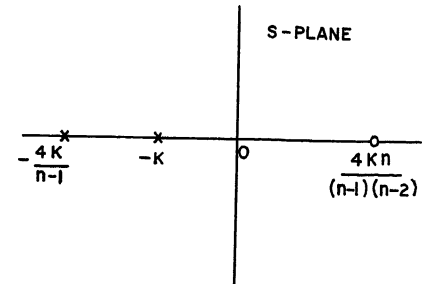
Fig. 8. Pole-zero plot of θ_0/θ_i (s) of system in Fig. 7
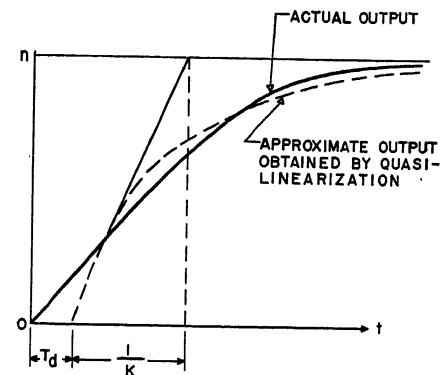
Fig. 9. Actual and approximate output of nonlinear system in Fig. 2

comparing the Maclaurin's series expansion of the ratio of the Laplace transforms of the arbitrary input and output with a similar expansion of the algebraic transfer function to be determined. Equating the coefficients of lower order terms of s in these two series yields the parameters of the approximate algebraic transfer function. Matching the lower order terms of s is a sensible basis of approximation in the present case because the approximate transfer function is expected to be followed by other transfer functions of the low-pass nature in practical systems.

As an illustration to the Pade's approximation, the transcendental transfer function $G_a(s)$ in equation 1 is to be approximated by an algebraic transfer function $\tilde{G}_a(s)$ of the first order. Expanding equation 1 into a Maclaurin's series

$$G_a(s) = \frac{(n-1)^2}{2nK}s - \frac{(n-1)^3}{3nK^2}s^2 + \frac{(n-1)^4}{4nK^3}s^3 - \cdots$$

(3)

Fig. 10. Nonlinearized servomechanism

Fig. 12. Linearization of nonlinear servo in its high-gain region

A—Block diagram
B—Frequency asymptotes for $K\alpha > 1/T$
C—Frequency asymptotes for $K\alpha < 1/T$
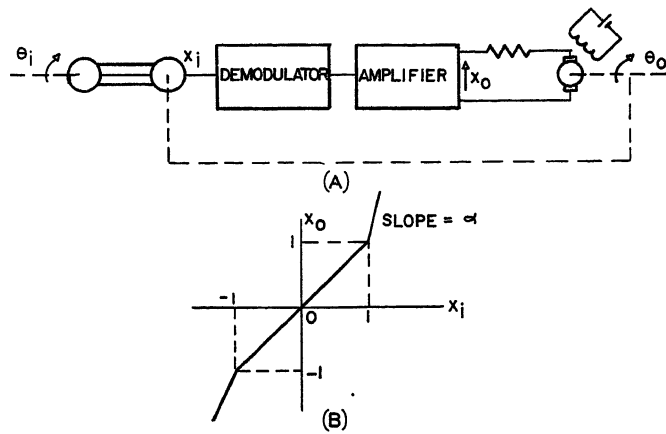
Expanding the algebraic transfer function into a Maclaurin's series

$$\tilde{G}_a(s) = \frac{a_0 + a_1 s}{b_0 + b_1 s}$$

$$= (a_0 + a_1 s)\frac{1}{b_0}\left[1 - \frac{b_1}{b_0}s + \left(\frac{b_1}{b_0}\right)^2 s^2 - \left(\frac{b_1}{b_0}\right)^3 s^3 + \ldots\right] \quad (4)$$

where the parameters a_0, a_1, b_0 and b_1 are to be determined. Equating the co-efficients of s^0, s^1, and s^2 terms in equations 3 and 4 gives

$$\frac{a_0}{b_0} = 0 \quad (5A)$$

$$-\frac{a_0 b_1}{b_0^2} + \frac{a_1}{b_0} = \frac{(n-1)^2}{2!nK} \quad (5B)$$

$$\frac{a_0 b_1^2}{b_0^3} - \frac{a_1 b_1}{b_0^2} = \frac{-(n-1)^3}{3!nK^2} \quad (5C)$$

Any one of the four parameters a_0, a_1, b_0, and b_1 may be set to an arbitrary nonzero number without affecting $\tilde{G}_a(s)$. The usual choice is to equate b_0 to unity. Then the three simultaneous equations 5 will uniquely determine $\tilde{G}_a(s)$. Equating coefficients of s terms in equations 3 and 4 higher than the second order would result in contradiction. This means $\tilde{G}_a(s)$ approximates $G_a(s)$ to the second degree. For better approximation, $\tilde{G}_a(s)$ must contain more terms. However, for ordinary design work, $\tilde{G}_a(s)$ usually need not contain terms higher than the second order. For rough estimates of transient characteristics, even a second-degree approximation as in equation 4 is sufficient in most cases.

In quasi-linearization, the given arbitrary output is in general a finite pulse. Usually the part of the system transient which follows the termination of the pulse is of interest. When this is the case, and when $\tilde{G}_a(s)$ contains but a few terms, modification of the Pade's approximation

in obtaining $\tilde{G}_a(s)$ will lead to more satisfactory results. For example, suppose that a proper algebraic transfer function $\tilde{G}_a(s)$ is to be found, which has a step response approximating the trapazoidal pulse $a(t)$ given in Fig. 6. If the Pade's approximation is used in the same manner as in setting up equations 3 to 5, the following expression of $\tilde{G}_a(s)$ will be obtained.

$$\tilde{G}_a(s) = \frac{(m/n)(T_d + T_r/2)s}{\left(\dfrac{3T_d^2 + T_r^2 + 3T_d T_r}{6T_d + 3T_r}\right)s + 1} \quad (6)$$

where n is the magnitude of the input step, and m, T_d and T_r are given in Fig. 6. The actual step response corresponding to $\tilde{G}_a(s)$ is a single time-constant exponential shown in the same figure and denoted by $\tilde{a}(t)$. The trapazoidal pulse $a(t)$ is approximated by $\tilde{a}(t)$ to the second degree in the sense that they have the same areas and the same first moments. That is

$$\int_0^\infty \tilde{a}(t)dt = \int_0^\infty a(t)dt \quad (7A)$$

$$\int_0^\infty t\tilde{a}(t)dt = \int_0^\infty ta(t)dt \quad (7B)$$
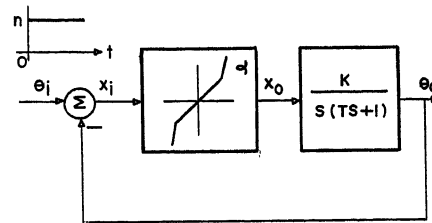
Fig. 11. Block diagram of normalized system in Fig. 10
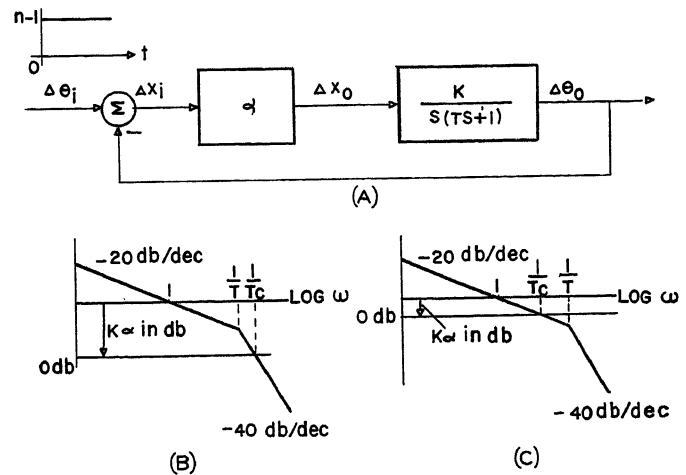
However, the time constant of $\tilde{a}(t)$ is so long that the magnitude of $\tilde{a}(t)$ is still considerable, long after the pulse $a(t)$ terminates. This will cause significant error in evaluating the system transient shortly after the termination of $a(t)$. Thus, a better expression of a first-order $\tilde{G}_a(s)$ is

$$\tilde{G}_a(s) = \frac{\dfrac{m}{n}\left(T_d + \dfrac{T_r}{2}\right)s}{\dfrac{1}{4}(T_d + T_r)s + 1} \quad (8)$$

Now, $\tilde{a}(t)$ corresponding to $\tilde{G}_a(s)$ in equation 8 retains the general shape as $\tilde{a}(t)$ in Fig. 6 and still satisfies equation 7(A), and its time constant is so short that it decays to almost zero when $a(t)$ terminates at $t = T_d + T_r$.

Integration of Approximate Methods with New Quasi-Linearization Technique

The integration of the approximate methods with the new quasi-linearization technique can best be explained by carrying on the transient study in the previous example. The fictitious transfer function $G_a(s)$ in Fig. 5 may now be replaced by the approximate algebraic transfer function $\tilde{G}_a(s)$ given by equation 8 with the substitution that $m = n-1$, $T_d = 0$ and $T_r = (n-1)/K$. This results in the block diagram in Fig. 7. The over-all transfer function of the system is then

$$\frac{\theta_0}{\theta_i}(s) = \frac{K}{s+K} \cdot \frac{\dfrac{-(n-2)(n-1)}{4Kn}s + 1}{\dfrac{n-1}{4K}s + 1} \quad (9)$$
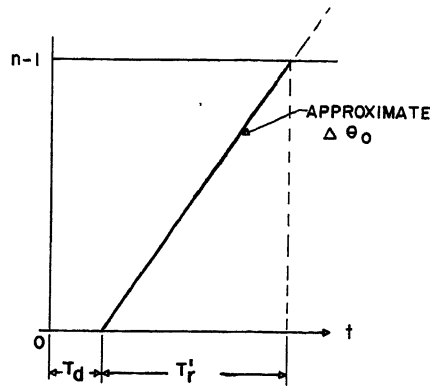
Fig. 13. Estimated output in initial transient stage
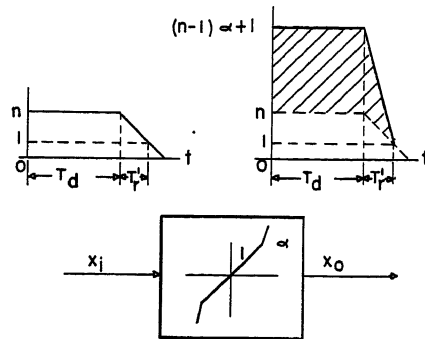
Fig. 14. Testing nonlinear element by approximate x_i

(A)

LINEARIZED SYSTEM IN THE LOW GAIN REGION   EQUIVALENT EFFECT OF NONLINEARITY

(B)
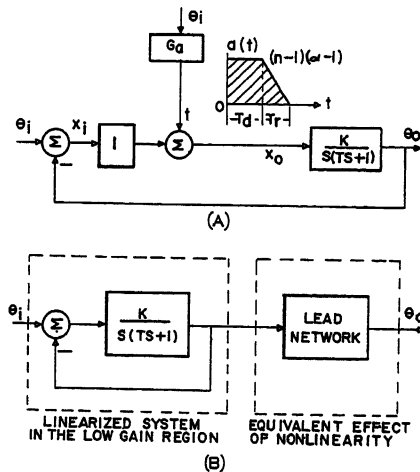
Fig. 15. Quasi-linearized system of Fig. 11

Now, the approximate methods developed in the Appendix can be used to study the quasi-linearized nonlinear system as if it were linear. The pole-zero configuration of the transfer function of equation 9 is displayed in Fig. 8 (poles designated by crosses, zeros by circles). The system response is monotonic as there are no complex poles. The pole closer to the origin determines the rise time, and the approximate delay time is given by $T_d = (n-1)/8K + (n-1)(n-2)/(8Kn)$; see the Appendix. Thus the approximate output may be easily sketched, as shown in Fig. 9.

The pole-zero plot in Fig. 8 offers to the designer a clear perspective of the relationship between system response and system parameters. The normalized system in Fig. 2 may be considered to represent an actual system having a limiter of ceiling equal to $1/n$ and a unit step input. Reducing the ceiling, or increasing n, shifts the pole at $-(4K)/(n-1)$ and the zero at $(4Kn)/(n-1)$ $(n-2)$, for $n>2$, closer to the origin. This corresponds to an increase in delay time and therefore a slower response. When the ceiling is so low or n is so large that $(4K)/(n-1)$ becomes smaller than K, the rise time becomes dominated by the pole at $-(4K)/(n-1)$ and the system speed of response is greatly reduced. On the other hand, raising the ceiling or decreasing n shifts the pole and the zero away from the origin, thereby resulting in a shorter delay time and faster response. For $1<n<2$, the zero is in the left-half plane and effects further reduction in delay time. The zero never falls between the two poles since $|n/(n-2)|>1$ and $4/(n-1)>1$ for $1<n<2$. Finally, as the ceiling approaches the magnitude of the input, or as n approaches unity, both the pole and the zero move to infinity. The system then operates in the unlimited region during the entire transient, and the transfer function of equation 9 becomes simply $(K)/(s+K)$, which is the transfer function of the linearized system. The case of $n<1$ has no meaning for equation 9 as this condition contradicts with the original assumption for the development of the equation.

These results check with physical reasoning and intuition. The transient study also points out that as the limiter ceiling is raised beyond a certain limit, the rise time is governed by the time constant $1/K$, and further elevation of the ceiling is not an efficient way to speed up the system response.

Transient Study of Nonlinearized Servomechanism

Application of the proposed quasi-linearization technique to more complicated transient problems is illustrated in the following example. Consider a typical positional servomechanism in Fig. 10(A). The gain of the demodulator-amplifier is deliberately nonlinearized to the form shown in Fig. 10(B). This "inverted saturation" or "hard-spring" effect has the advantage that the system responds quickly to large inputs, yet possesses considerable stability, and the bandwidth of system response for small-noise signals is minimized. Assume that
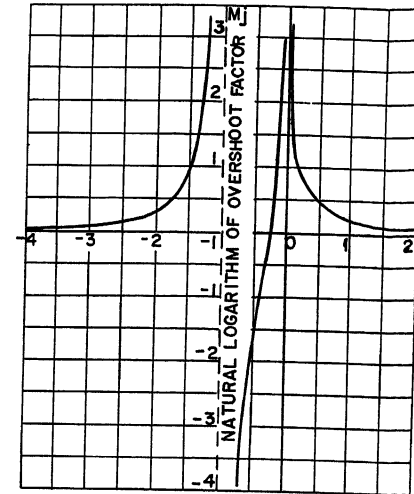
Fig. 16. Overshoot-factor chart for double-order dominant real pole at -1
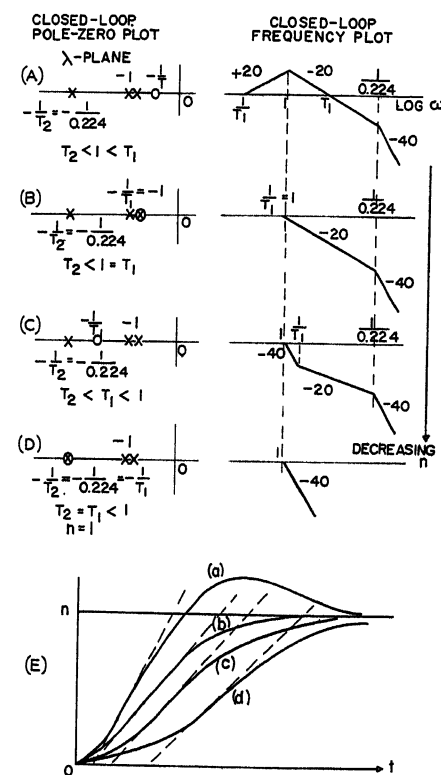
Fig. 17. Effect of varying input magnitude n

the amplifier has negligible time constant and the motor operates in the linear range during all transients. The block diagram of the system is shown in Fig. 11 with the nonlinear gain normalized. The system is subjected to a step input. The effect of varying the step input magnitude n, $(n>1)$, and of varying the gain α upon the system transient response will be studied. Insertion of linear lead and lag networks in the nonlinear system will be considered.

The same system for T and K chosen to give critical damping in the low-gain region has been analyzed by Blumenthal
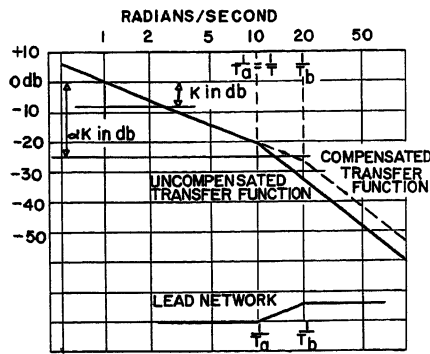
Fig. 18. Design of lead network compensation

and Beck, who solved the piecewise linear differential equations by the ordinary step-by-step method.[8] Their results will provide a convenient check for results in the present example.

Since $n>1$, the nonlinear element will operate in the high-gain region immediately following the initiation of the step input at $t=0$. To quasi-linearize the system, the approximate input to the nonlinear element, x_i, during the system's operation in the high-gain region, is first to be found. Linearization of the system in the high-gain region results in the block diagram shown in Fig. 12(A). The incremental quantities are related to their corresponding total quantities in the following manner.

$$\theta_i = 1 + \Delta\theta_i \tag{10A}$$

$$\theta_0 = 0 + \Delta\theta_0 \tag{10B}$$

$$x_i = 1 + \Delta x_i \tag{10C}$$

$$x_0 = 1 + \Delta x_0 \tag{10D}$$

The frequency asymptotes of the linearized systems are shown in Fig. 12(B) and 12(C), depending on the relative magnitude of $K\alpha$ with respect to T. $1/T_c$ is the asymptote crossover frequency and may be expressed as

$$T_c = \sqrt{\frac{T}{K\alpha}} \quad \text{for} \quad K\alpha \geq \frac{1}{T} \tag{11A}$$
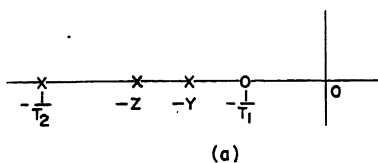
$$T_c = \frac{1}{K\alpha} \quad \text{for} \quad K\alpha \leq \frac{1}{T} \tag{11B}$$

According to the rules of thumb developed in the Appendix, the initial transient stage of the output θ_0 may be approximated by two straight-line segments, as in Fig. 13, where

$$T_d = \frac{1}{2}T_c = \frac{1}{2}\sqrt{\frac{T}{K\alpha}} \quad \text{for} \quad K\alpha \geq \frac{1}{T} \quad \tag{12A}$$

$$T_d = \frac{1}{2}T \quad \text{for} \quad K\alpha \leq \frac{1}{T} \tag{12B}$$

$$T_r' = \frac{3}{2}T_c = \frac{3}{2}\sqrt{\frac{T}{K\alpha}} \quad \text{for} \quad K\alpha \geq \frac{1}{T} \tag{13A}$$

$$T_r' = \frac{3}{2}T_c = \frac{3}{2}\frac{1}{K\alpha} \quad \text{for} \quad K\alpha \leq \frac{1}{T} \tag{13B}$$

The approximate x_i may be readily derived from the approximate $\Delta\theta_0$ and is used to test the nonlinear element, as shown in Fig. 14, in which the shaded area in the output wave form indicates the effect of the existence of a high-gain region. Following the procedure outlined in the previous example, the quasi-linearized system in Fig. 15(A) may be deduced. Using equation 8 to approximate $G_a(s)$ and by proper substitution and block diagram manipulation, the resultant block diagram of Fig. 15(B), in which the lead network has the following transfer function, may be obtained.

$$\frac{T_1 s + 1}{T_2 s + 1} \tag{14}$$

where

$$T_1 = \frac{1}{2}\sqrt{\frac{T}{K\alpha}}\left[1 + \frac{n-1}{n}(\alpha-1)\frac{5}{2}\right]$$
$$\text{for} \quad K\alpha \geq \frac{1}{T} \tag{15A}$$

$$T_1 = \frac{1}{8}\left(T + \frac{3}{K\alpha}\right)\left[1 + \frac{n-1}{n}(\alpha-1)\times\right.$$
$$\left.\frac{2(2TK\alpha+3)}{TK\alpha+3}\right] \quad \text{for} \quad K\alpha \leq \frac{1}{T} \tag{15B}$$

$$T_2 = \frac{1}{2}\sqrt{\frac{T}{k\alpha}} \quad \text{for} \quad K\alpha \geq \frac{1}{T} \tag{15C}$$

$$T_2 = \frac{1}{8}\left(T + \frac{3}{K\alpha}\right) \quad \text{for} \quad K\alpha \leq \frac{1}{T} \tag{15D}$$

Note that although the effect of nonlinearity upon transient response is equivalent to that of putting a linear lead network outside the loop, the advantage of feeding back the actual output signal to the input for minimizing drift is not retained if a strictly linear system is constructed according to Fig. 15(B).

To illustrate the use of the results given in the Appendix, assume that $n=5$, $\alpha=5$, and $TK=1/4$. Since $\alpha K>1/T$ in this case, the delay time T_d and the rise time T_r of the output θ_0 are given by equations 12 (A) and 13 (A).

$$T_d = \frac{1}{2}\sqrt{\frac{T}{K\alpha}} = 0.224\sqrt{\frac{T}{K}} \tag{16}$$

$$T_r = \frac{n}{n-1}T_r' = \frac{n}{n-1}\frac{3}{2}\sqrt{\frac{T}{K\alpha}} = 0.839\sqrt{\frac{T}{K}} \tag{17}$$

The over-all transfer function is

$$\frac{\theta_0}{\theta_9}(s) = \frac{1}{(\lambda+1)^2}\frac{0.224\times9\lambda+1}{0.224\lambda+1} \tag{18}$$

where $\lambda = \sqrt{T/K}\,s$. The double pole at $\lambda = -1$ is evidently the dominant one. The overshoot factor is contributed mainly by the zero at $-(1)/(0.224\times9)$ and is read from Fig. 16 to be -1.9 in natural logarithm or 0.15 per unit, which means that the overshoot value is $5\times(1+0.15) = 5.75$. The time at which the overshoot occurs is given by equation 35 to be 2.28 in the time scale corresponding to λ. The results all check within 10 per cent with the exact solution obtained by Blumenthal and Beck.[8]

Next will be considered the effect of varying n on the transient response.
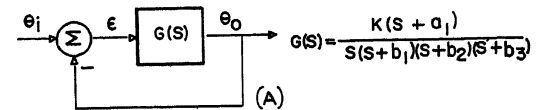
Fig. 19 (left). Nonlinear servo with lead network compensation

A—Pole-zero plot for $T_b < 0.1$ but still relatively large

B—Transient responses

(A)

$$G(S) = \frac{K(S+a_1)}{S(S+b_1)(S+b_2)(S+b_3)}$$

(B)
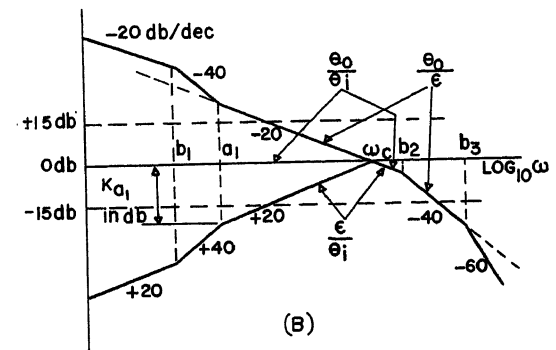
Fig. 20 (right). Frequency response of typical linear feedback system
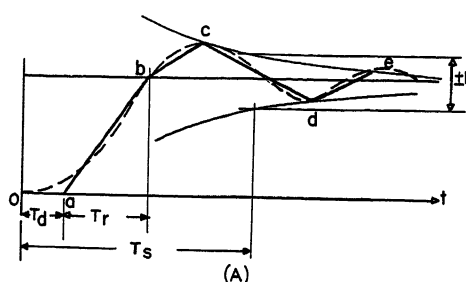
Fig. 21 (left). Transient characteristics of oscillatory and monotonic responses
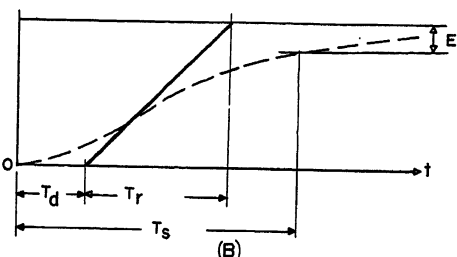
Fig. 22 (right). Corresponding ranges in time domain, frequency plot, and complex frequency domain

A—Time response
B—Open-loop frequency plot
C—Closed-loop poles and zeros

Again assume $\alpha=5$ and $TK=1/4$. Since equations 15 indicate that T_2 is independent of n, the pole-zero plot and the frequency-response plot of θ_0/θ_i may be sketched for various magnitudes of n as shown in Fig. 17. The corresponding transient responses are sketched in the same figure by using the rules of thumb described in the Appendix. Increasing n in the normalized nonlinear system is equivalent to narrowing the low-gain range in an actual system with constant input magnitude. Thus, the apparent effect of introducing the high-gain region into the system which is critically damped in the low-gain region is to speed up its response. However, if the low-gain region is too narrow, the overshoot may be excessive.

Varying the magnitude of the high gain α will affect T_1 and T_2. The θ_0/θ_i plots for various magnitudes of α and their corresponding transient responses may be sketched in a like manner. The effect of decreasing α is similar to that of decreasing n.

The insertion of a linear compensation network inside the feedback loop affects the quasi-linear block diagram of Fig. 15 in two respects. First, it affects the linearized system in the low-gain region. In this respect, the relative position of the compensation network in the loop with respect to the nonlinear element is immaterial. Second, the compensation network affects the test function for the nonlinear element and therefore the equivalent lead network outside the loop. In this respect, the question whether the compensation network precedes or follows the nonlinear element has significance. We shall consider the effect of inserting a lead network $G_c(s)=(T_a s+1)/(T_b s+1)$, $T_a>T_b$, between the nonlinear device and the motor. Note that the transfer func-

tion $G_c(s)K/s(Ts+1)$ still has low-pass nature, and hence the Pade's approximation may still be used in evaluating $G_a(s)$.

To fix ideas, assume that $n=5$, $\alpha=7$, $T=0.1$, and $K=2.5$. The system with this set of parameters is found to exhibit an excessive overshoot in its step response. A lead network $G_c(s)$ is to be designed to suppress the overshoot and speed up the system response, if possible. The open-loop frequency asymptotes of the system is shown in Fig. 18. That the bandwidth of the system in the low-gain region be unchanged for noise effect considerations, the lead network should have its break frequencies above 2.5 radians per second. For convenience, make the zero (at $-1/T_a$) of the lead network coincide with the open-loop pole at $-1/T$. This will extend the -20 decibels (db) per decade (dec) asymptote to the break point at $1/T_b$. The rise time T_r' is given by 3/2 divided by ω_c, where $\omega_c=17.5$ is the frequency at the intersection of the extended asymptote with the $\alpha k-$db line. The delay time is given by $(1/2)T_b$. The transfer function $\bar{G}_a(s)$ defined in Fig. 15 may therefore be approximated, according to equation 8 by

$$\bar{G}_a(s) = \frac{\frac{(n-1)(\alpha-1)}{n}\left(\frac{T_b}{2}+\frac{1}{2}\times\frac{3}{2}\times\frac{1}{17.5}\right)s}{\frac{1}{4}\left(\frac{T_b}{2}+\frac{3}{2}\times\frac{1}{17.5}\right)s+1}$$

$$= \frac{(2.4T_b+0.206)s}{(0.125T_b+0.0214)s+1} \quad (19)$$

The over-all transfer function is

$$\frac{\theta_0}{\theta_i}(s)=\frac{1}{\left(\frac{s}{y}+1\right)\left(\frac{s}{z}+1\right)}[1+\bar{G}_a(s)]$$
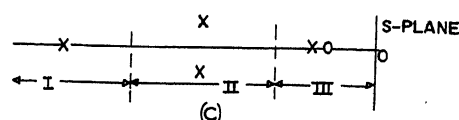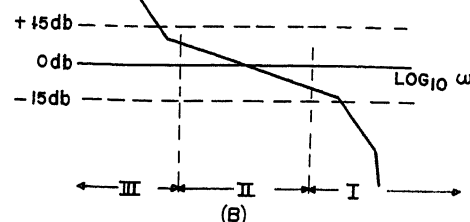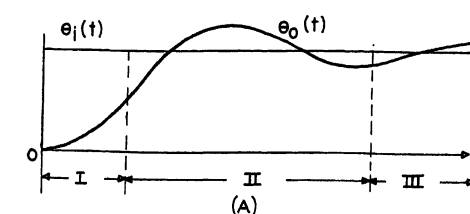
$$= \frac{1}{\left(\frac{s}{y}+1\right)\left(\frac{s}{z}+1\right)}\frac{T_1 s+1}{T_2 s+1} \quad (20)$$

where

$$y=\frac{1-\sqrt{1-10T_b}}{2T_b} \quad (21A)$$

$$z=\frac{1+\sqrt{1-10T_b}}{2T_b} \quad (21B)$$

$$T_1=2.53T_b+0.227 \quad (21C)$$

$$T_2=0.125T_b+0.0214 \quad (21D)$$

The poles at $-y$ and $-z$ are those of the linearized transfer function for the low-gain region. The pole at $-1/T_2$ and the zero at $-1/T_1$ are due to the nonlinearity. The pole-zero plot of θ_0/θ_i shown in Fig. 19(A) corresponds to the case of a relatively large T_b ($T_b<0.1$ in all cases). The system response will have an overshoot and the rise time is mainly determined by y. On the other hand, if T_b is small, the zero at $-1/T_1$ will move to the left of $-y$. Then there will be no over-shoot in the transient response and the delay time will become negative, but the rise time is still governed by y. This means a rapid inital rise of the system to a fraction of the steady-state value followed by a transient which has a long tail. The most desirable choice for T_b is such that the zero at $-1/T_1$ cancels the pole at $-y$. Then there will be negligible overshoot and the speed of response is determined by the pole at $-z$, resulting in a much faster system. The proper value of T_b is obtained by equating y to $1/T_1$, which are given by equations 21(A) and 21(C) respectively. This yields $T_b=0.047$. Hence the desired lead network is $G_c(s)=$ $(0.1s+1)/(0.047s+1)$. The transient re-

ponses for various values of T_b are sketched in Fig. 19(B). The results have been checked by an analogue computer.

A similar approach may be used to design a lag network in order to suppress overshoot. However, it was found that when the lag network was introduced substantially below the crossover frequency, the optimum speed of response was much lower than that obtainable by a lead network compensation. When the lag network was introduced near the crossover frequency, system instability resulted, as could be predicted by a describing function analysis.[6]

Conclusions

Nonlinear differential equations have no general solutions. The study of nonlinear system dynamics therefore requires some form of approximations and simplifications. The novel quasi-linearization technique developed in this paper is one of these reasonable approximations. Its value as a tool of synthesis has been emphasized and demonstrated.

Although the proposed method is straightforward when the nonlinearity considered is relatively simple, the complexity in applying the method increases rapidly when there are several nonlinear elements in the system and when the system oscillates several times between two piecewise linear regions. It is believed that the design of a certain class of nonlinear systems can be carried out most efficiently by first using the describing function method to insure stability, then applying the present method to study the transient problem, and finally using a computer or performing actual tests on the designed prototype to check test results with the design objectives. Cut-and-try procedure is inevitable in all design problems which have no unique solutions. Intelligent guidance to nonlinear system design is the important function provided by the quasi-linearization techniques.

Finally, it should be pointed out that the approximate methods described in this paper for studying transient response of linear systems and for deriving an algebraic transfer function with arbitrary input-output relationship are not the only means to accomplish their ends. The author wishes to bring forth in this paper the concept of a new application of quasi-linearization and not to propose a stereotyped quasi-linearization procedure. If the paper encourages further development of quasi-linearization techniques, the author's goal in presenting this paper shall have been achieved.

Appendix. Quick Evaluation of Approximate Transient Response of Linear Systems

Evaluating Closed-Loop Poles

Approximate locations of closed-loop poles can be quickly determined by combining a well-known approximate method which has been thoroughly treated by Biernson[9] and the feedback charts developed by Yeh.[10]

Consider the linear closed-loop system and its open-loop gain-frequency asymptote plot in Fig. 20. When $|G| \geq 15$ db, the following approximation may be written

$$\frac{\epsilon}{\theta_i} = \frac{1}{1+G} \approx \frac{1}{G} \qquad (22)$$

When $|G| \leq -15$ db, the following approximation may be written

$$\frac{\theta_0}{\theta_i} = \frac{G}{1+G} \approx G \qquad (23)$$

Hence, the following conclusion may be drawn: When $|G| \geq 15$ db, the zeros of the open-loop transfer function are the approximate closed-loop poles, and when $|G| \leq -15$ db, the poles of the open-loop transfer function are the approximate closed-loop poles. The choice of ± 15 db as the upper and lower limits is arbitrary within the bounds of engineering judgment. Thus, two closed-loop poles of the system in Fig. 20 are approximately at $s = -a_1$, and $s = -b_3$.

For -15 db $\leq |G| \leq +15$ db, the break frequencies outside the ± 15-db band may be ignored. The frequency plot in Fig. 20 outside the ± 15-db band is then approximated by the slant dashed lines which are extensions of the actual asymptotes crossing the ± 15-db lines. In other words, for $|G|$ within the ± 15-db limits, it may be considered that $b_1 < a_1 \ll |s| \ll b_3$, and hence the following approximations may be taken.

$$G(s) = \frac{K(s+a_1)}{(s+b_1)(s+b_2)(s+b_3)} \approx \frac{K/b_3}{s(s+b_2)} \qquad (24)$$

$$\frac{\theta_0}{\theta_i}(s) = \frac{G(s)}{1+G(s)} \approx \frac{K/b_3}{s(s+b_2)+K/b_3} \qquad (25)$$

The other two approximate closed-loop poles of the system may then be easily evaluated from equation 25 by solving a quadratic equation.

It is found that in almost all practical systems, the $G(s)$ within ± 15-db limits may be approximated by a transfer function not higher than the third order. When $G(s)$ is of the third order, the closed-loop poles may best be located not by solving the third-order characteristic equation, but by using the feedback charts developed by Yeh.[10] Yeh's feedback charts, which were derived as general root loci, contain families of root loci for the closed-loop poles with the loop gain and the relative positions of the open-loop poles and zeros as the varying parameters. These charts are valuable not only in determining the closed-loop poles but in helping the designer to choose proper open-loop functions to shift the closed-loop poles to desired locations.

Meaning of Transient Characteristics

In general, the designer of a control system is interested only in the prominent features of the transient response. Thus, the actual transient response indicated by the dashed curves in Fig. 21(A) may be represented by the solid straight-line segments $Oabcde$ in the same figure without losing its significant characteristics. The delay time T_d, the rise time T_r, the magnitude of overshoots and undershoots and the time at which they occur, and the settling time T_s characterize the complete transient response. Notice that while T_s may be defined as the time after which the envelope of the error becomes smaller than a predetermined value E, there are no clear-cut definitions for T_d and T_r.

When the response is monotonic as shown in Fig. 21 (B), it may be shown mathematically that the response may be approximated by a fixed time delay followed by a single exponential rise.[11] In this case, T_d has a convenient definition, and the rise time T_r should be a measure of the time constant of the single exponential rise.

Division of Transient Response into Three Ranges

The closed-loop transient response in the t-domain may be subdivided into three ranges, each corresponding to a certain range in the open-loop frequency response and in the closed-loop pole-zero plot as in Fig. 22. This is based on the theory that a function near infinity of the t-domain corresponds to its transform function near the origin of the s-domain, and vice versa. Range I is the subtransient stage in the time domain, in which the transient terms corresponding to the poles remote from the origin of the s-plane or those above the crossover frequency in the frequency plot play the important role. Range II in the t-domain is the main transient stage which is characterized by the rise time and the maximum overshoot if there is one. This range corresponds to that near the crossover frequency in the frequency plot. The transient term which has the largest magnitude among others and has a relatively long time constant dominates in this range. Range III is mainly the tail of the transient response, which is affected by the transient terms corresponding to open-loop zeros above the +15-db line in the frequency plot if there is any. Note that, in Fig. 22, the real pole nearest the origin does not correspond to the dominant transient term because of its proximity to a zero.

Evaluating Delay Time

Delay time is contributed mainly by the poles and zeros in range I. Consider a transfer function which has a pole $-\beta$ and a zero $-\alpha$ quite remote from the origin $s=0$. The transfer function may be written as

$$\frac{\theta_0}{\theta_i}(s) = F(s)\frac{\left(1+\frac{s}{\alpha}\right)}{\left(1+\frac{s}{\beta}\right)} \qquad (26)$$
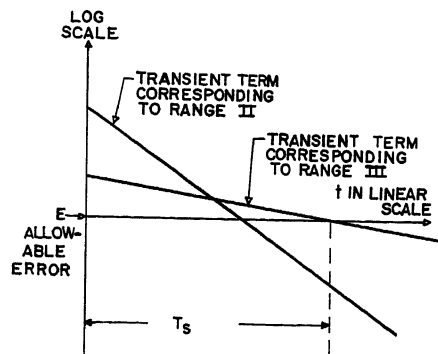
Fig. 23. Determination of settling time

Now

$$\epsilon^{s/\alpha} = 1 + \frac{s}{\alpha} + \frac{s^2}{2!\alpha^2} + \frac{s^3}{3!\alpha^3} + \cdots \qquad (27A)$$

$$\epsilon^{-s/\beta} = 1/\epsilon^{s/\beta} = 1 / \left(1 + \frac{s}{\beta} + \frac{s^2}{2!\beta^2} + \frac{s^3}{3!\beta^3} + \cdots \right) \qquad (27B)$$

For large α and β and small s, the s terms in equations 27(A) and 27(B) higher than the first order may be neglected and equation 26 may be written as

$$\frac{\theta_0}{\theta_i}(s) \approx F(s)\epsilon^{+s/\alpha}\epsilon^{-s/\beta} \qquad (28)$$

Thus, so far as the transient response for $t \gg 0$ (corresponding to s being small) is concerned, the effect of the pole $-\beta$ and the zero $-\alpha$ is in their contribution to a delay time $T_d = 1/\beta - 1/\alpha$. If the zero were at $+\alpha$; i.e., in the right half of the s-plane, the delay time would be $T_d = 1/\beta + 1/\alpha$. In practical cases, however, the poles and zeros in range I are not quite remote from the origin. The rule of thumb for evaluating their contribution to the delay time is to apply a factor of $1/2$ to their time constants.

The delay time attributed to a pair of complex poles is equal to $1/2\omega_n$, where ω_n is the undamped natural frequency, regardless of its damping factor. This rule was established through a pictorial study of the transient response of a second-order oscillatory system. ω_n of the closed-loop complex poles is approximately equal to ω_c, the crossover frequency of the open-loop frequency plot, if the crossover asymptote has a slope of -40 db per dec, and is approximately equal to the first break frequency beyond ω_c if the crossover asymptote has a slope of -20 db per dec.

According to these rules of thumb, the delay time associated with the system of Fig. 20 is approximately equal to $1/b_2 + 1/b_3$. ω_c contributes nil to the delay time since the crossover asymptote has a slope of -20 db per dec.

Evaluating Rise Time

If a system response is oscillatory, its dominant transient term may be approximated by that of a second-order system described by the differential equation

$$\frac{1}{\omega_n^2}\frac{d^2\theta_0}{dt^2} + \frac{2\zeta}{\omega_n}\frac{d\theta_0}{dt} + \theta_0 = \theta_i \qquad (29)$$

The rising portion of the response may

be approximated by a straight-line tangent to the response time-function through the inflection point at $d^2\theta_0/dt^2 = 0$. Thus the rise time is related to $2\zeta/\omega_n$, where ω_n is in the neighborhood of the open-loop crossover frequency ω_c. The rule of thumb, for lack of a better method, is to take the rise time as approximately equal to $3/2\omega_c$ in practically all cases of oscillatory response.

In the case of a monotonic system response, the dominant transient term is a single exponential with a time constant approximately equal to $1/\omega_c$. Then $3/2\omega_c$ is the time for the response to rise to about 77 per cent of its steady-state value.

Evaluating Overshoots and Undershoots

A method of quickly estimating the magnitude of overshoots and undershoots as well as the time at which they occur from the given closed-loop pole-zero configuration has been developed by Mulligan.[12] Results were given in the form of charts for the case of a pair of dominant complex poles. The method is based on the evaluation of the overshoot contributed by the dominant transient term only. The approximation lies in the negligence of all transient terms other than the dominant one, but the presence of poles and zeros other than the dominant pair is not ignored in determining the magnitude of the dominant transient term. Extension of Mulligan's method to the case of a dominant double-order real pole is given in the following.

Consider a system which has m real zeros and n real poles and that the system function has been normalized so that the dominant double-order pole is at -1. Then the Laplace transform of the amplitude-normalized step response is

$$F(s) = \frac{\overset{n}{\Pi}\delta_n}{\overset{m}{\Pi}\gamma_m} \frac{\overset{m}{\Pi}(s+\gamma_m)}{s(s+1)^2\overset{n}{\Pi}(s+\delta_n)} \qquad (30)$$

where γ_m and δ_n indicate locations of zeros and poles respectively. The time response may be written as

$$f(t) = 1 + A_1 t \epsilon^{-t} + A_2 \epsilon^{-t} + \text{other}$$
$$\text{nondominant transient terms} \qquad (31)$$

where

$$A_1 = -\frac{\overset{m}{\Pi}\dfrac{-1+\gamma_m}{\gamma_m}}{\overset{n}{\Pi}\dfrac{-1+\delta_n}{\delta_n}} \qquad (32A)$$

$$A_2 = -\frac{\overset{m}{\sum}\left[\dfrac{1}{\gamma_m}\overset{m-1}{\Pi}\dfrac{-1+\gamma_{m-1}}{\gamma_{m-1}}\right]}{\overset{n}{\Pi}\dfrac{-1+\delta_n}{\delta_n}} +$$

$$\frac{\overset{m}{\Pi}\dfrac{-1+\gamma_m}{\gamma_m}}{\overset{n}{\Pi}\dfrac{-1+\delta_n}{\delta_n}}\overset{n}{\sum}\dfrac{1}{-1+\delta_n} -$$

$$\frac{\overset{m}{\Pi}\dfrac{-1+\gamma_m}{\gamma_m}}{\overset{n}{\Pi}\dfrac{-1+\delta_n}{\delta_n}} \qquad (32B)$$

If the nondominant transient terms are neglected, the derivative of time response is

$$f'(t) \approx \epsilon^{-t}[A_1(1-t) - A_2] \qquad (33)$$

The zeros of $f'(t)$ or the maxima and minima of $f(t)$ occur at

$$t_m \approx 1 - \frac{A_2}{A_1} \qquad (34)$$

Substituting equations 32(A) and 32(B) into equation 34

$$t_m = -\overset{m}{\sum}\frac{1}{-1+\gamma_m} + \overset{n}{\sum}\frac{1}{-1+\delta_n} \qquad (35)$$

Substituting equations 32(A), 32(B) and 35 into equation 31, one may evaluate the overshoot factor in per units as follows

$$[f(t_m) - 1] \approx -\frac{\overset{m}{\Pi}\dfrac{-1+\gamma_m}{\gamma_m}}{\overset{n}{\Pi}\dfrac{-1+\delta_n}{\delta_n}}\epsilon^{\overset{m}{\sum}\frac{1}{-1+\gamma_m} - \overset{n}{\sum}\frac{1}{-1+\delta_n}} \qquad (36)$$

From equation 36 it is seen that the overshoot factor in natural logarithm contributed by the jth real zero is

$$M_j = \ln\frac{D_{j1}}{D_{j0}} + \frac{1}{D_{j1}} \qquad (37)$$

where $D_{j1} = -1+\gamma_m$ and $D_{j0} = \gamma_m$ are the distances from the zero to the double-order pole and to the origin respectively. Equation 37 applies to the overshoot factor due to a real pole if the signs are changed. The expression of equation 37 is plotted in Fig. 16 as an overshoot-factor chart. In view of equations 31 through 36, the following remarks may be made.

1. If the total number of poles and zeros of $F(s)/s$ between the origin and -1 is odd, there is an overshoot; if even, there is an undershoot, unless $(1 - A_2/A_1) < 0$.

2. If $(1 - A_2/A_1) < 0$, there is neither an overshoot nor an undershoot.

Likewise, an overshoot factor chart may be derived for the case of a pair of dominant real poles at -1 and $-\alpha$, $(\alpha < 1)$.[6] In this case, the following is pertinent: In most practical cases, if the total number of zeros and poles of the transfer function $F(s)/s$ between the origin and $-\alpha$ and that between the origin and -1 excluding the pole at $-\alpha$ are both odd, there is an overshoot; if both even, there is an undershoot; and if one number is odd and the other even, there is no overshoot or undershoot.

Evaluating Settling Time

The settling time is affected by transient terms corresponding to the poles in both ranges II and III. The relative significance of these terms depends on which term dominates the transient when the predetermined allowable value of error is neared. The magnitude of the transient term corresponding to range II may be determined from a modified form of Mulligan's charts.[12] The magnitude of the transient term corresponding to a pole in range III may be shown[6] to be approximately equal to the reciprocal of the open-loop gain at a frequency corresponding to

Chen—Techniques for Study of Nonlinear Feedback Control Systems

this pole in range III. Thus, for the system of Fig. 20 the magnitude of the transient term corresponding to the closed-loop pole approximately at $-a_1$ is equal to k_{a1}; see Fig. 20. Each of these transient terms is then plotted on semilogarithmic graph paper as in Fig. 23. The transient term which reaches the allowable error last determines the settling time.

References

1. A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms, Ralph J. Kochenburger. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 270–84.

2. Sinusoidal Analysis of Feedback-Control Systems Containing Nonlinear Elements, E. Calvin Johnson. *AIEE Transactions*, vol. 71, pt. II, July 1952, pp. 169–81.

3. On Some Nonlinear Phenomena in Regulatory Systems (in Russian), L. C. Goldfarb. *Automatika i Telemekhanika*, vol. 8, no. 5, 1947, pp. 349–83. (English translation in National Bureau of Standards Report No. 1691, 1952.)

4. The Effects of Backlash and of Speed-Dependent Friction on the Stability of Closed-Loop Control Systems, A. Tustin. *Journal*, Institution of Electrical Engineers, London, England, vol. 94, pt. IIA, 1947, pp. 143–51.

5. The Analysis of Nonlinear Control Systems with Random Inputs, R. C. Booton, Jr. *Proceedings of the Symposium of Nonlinear Circuits Analysis*, Brooklyn Polytechnic Institute, Brooklyn, N. Y., Apr. 1953, pp. 369–91.

6. Quasi-Linearization Techniques for Transient Study of Nonlinear Feedback-Control Systems, K. Chen. *Thesis*, Massachusetts Institute of Technology, Cambridge, Mass., June 1954.

7. Contribution to the Theory of Servo Systems in the Time Domain (in Spanish), Garcia-Arbines y Calvo, A. *Revista de ciencia aplicada*, Madrid, Spain, pts. I, II, nos. 26, 27; May-June, July-August 1952, pp. 143–212, 303–24.

8. Transient Analysis of Nonlinearized Single-Lag Servomechanisms. J. S. Blumenthal, F. J. Beck. *Proceedings of the First U.S. National Congress of Applied Mechanics*, American Society of Mechanical Engineers, New York, N. Y., 1951, pp. 155–60.

9. Quick Methods for Evaluating the Closed-Loop Poles of Feedback Control Systems, George Biernson. *AIEE Transactions*, vol. 72, pt. II, May 1953, pp. 53–70.

10. The Study of Transients in Linear Feedback Systems by Conformal Mapping and the Generalized Root Locus Method, V.C.M. Yeh. *Thesis*, Massachusetts Institute of Technology, Cambridge, Mass., 1952.

11. Grundlagen der Selbsttatigen Regelung (in German), O. Schafer. Franzis-Verlag, Munich, Germany, 1953.

12. The Effect of Pole and Zero Locations on the Transient Response of Linear Dynamic Systems, J. H. Mulligan, Jr. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 37, 1949, pp. 516–29.

Discussion

Rudolf E. Kalman (Columbia University, New York, N. Y.): Dr. Chen's courageous and thoughtful effort toward a more comprehensive understanding of nonlinear dynamic systems containing feedback should be particularly welcome at this time when the practical solution of such problems is becoming increasingly more urgent. Unfortunately, progress in the nonlinear field during the past 10 years has been so slow that even at this time there is no broad agreement regarding the nature of the problems, the answers sought, or the strategy of attack. Recent work by the discusser[1,2] runs in many respects parallel to Dr. Chen's ideas as expressed in the paper. It is of interest to re-examine the fundamental problems in the light of these papers.

The principal task of nonlinear dynamic system analysis is to obtain a comprehensive picture, quantitative if possible, but at least qualitative, of what happens in the system if the variables are allowed or forced to move far away from their equilibrium positions or operating points. This is called global or in-the-large behavior. By contrast, it has long been recognized, in fact rigorously proved,[3] that if the system variables are perturbed only slightly from the operating points (local or in-the-small behavior), the problem may be considered linear, i.e., it can be linearized. This point is of very basic importance but may not be clear from the author's "A General Principle of Quasi-Linearization," discussed on page 355.

When one attempts to apply linearization to problems of global behavior, as for instance in the case of the describing-function method, an immediate difficulty arises in that the basic mathematical issues involved have not been resolved; indeed, not even thoroughly investigated. The describing-function method may be used to establish with confidence the magnitude and frequency of nearly sinusoidal oscillations in a nonlinear system when the fact that an oscillation of this type exists is known *a priori*, e.g., in the case of an oscillator. But it is also frequently implied in the current engineering literature that the describing-function method can be used to discover the existence of such oscillations. Nothing could be less consistent from a purely logical point of view. The very nature of the approximation involved presupposes the presence of nearly sinusoidal oscillations, and it would seem preposterous to use such a method, whose very validity is based on an assumption, to investigate the correctness of that assumption. Yet it is an empirical fact that the describing-function method gives correct answers in a number of practical situations (especially those where the nonlinearity is not very severe). For instance, the discusser's paper[2] contains an example where the conditions for the presence of an unstable limit cycle are derived by means of rigorous procedure; surprisingly enough, precisely the same conditions are obtained also by a far less satisfactory derivation based on the principles of the describing function. On the other hand, there are also many other examples, notably in cases where several nonlinearities are present, where the describing-function method appears utterly powerless to cope with the situation. The discusser regards the present status of the describing-function method as unsatisfactory; the ideas involved undoubtedly have merit but the range of their applicability and validity has not yet been convincingly established. In contrast to the describing-function method, Dr. Chen's approach to linearization does not require any *a priori* assumptions. As long as the piecewise linear approximation of the nonlinearity is accepted, one can always solve for a part of the transient response (corresponding to a straight-line segment in the approximation) exactly, since the transient then obeys a linear differential equation. This state of affairs was used in an earlier paper.[1] Since the problem for each straight-line segment of the nonlinearity is linear, the author is of course on solid grounds in applying and improving the usual methods of approximating linear transients.

Nevertheless, the paper fails completely in providing a satisfactory picture of global behavior of a nonlinear system. This is because the paper is basically a step-by-step procedure (although much more elegant and better integrated than the conventional numerical approach). Solutions can only be obtained after going through quite a few steps of approximations, equivalent representations, etc. Even after a complete equivalent representation is obtained, several additional steps are required to translate the results in the time domain and obtain actual approximate transient responses. The over-all picture is not available at any one time, but requires examination of a number of different solutions (say, transient responses to step inputs of varying magnitudes as in Fig. 17 of the paper) before the general state of affairs becomes apparent. This difficulty stems from the fact that the method is not correlated with the phase-plane (or phase-space) representation of nonlinear differential equations.

By contrast, consider the discusser's earlier published analysis of the problem of Fig. 11 of the paper, using phase-plane concepts.[1] After only a very slight amount of preparatory work which involves putting the differential equation in standard form and making some assumption about the relative magnitude of the gain corresponding to the various straight-line segments in the nonlinearity, the situation is sketched by inspection in the phase plane. It is then at once apparent that the transient will overshoot more and more times as the magnitude of the step inputs to the system is increased, but otherwise the system behaves in an entirely docile fashion. Almost as easily, quite similar conclusions may be obtained for ramp inputs which are not considered at all by Dr. Chen in the present study.

It should be strongly emphasized here that the phase-plane, or rather, phase space is not just a name for a clever method directed toward solving particular problems, but it is the most fundamental concept of the entire theory of ordinary differential equations (linear or nonlinear, time-dependent or time-invariant) at the present time.[4] This is because the phase-space is a way of representing (sometimes only in an abstract sense) all possible solutions of a differential equation. In linear systems, this is of possible but not very important; in nonlinear systems, it is absolutely essential,

because there is (up to the present, anyway) no other possible means of investigating global properties of solutions. There is a regrettable tendency at this time to try to treat difficult nonlinear problems with superficial generality, without clarifying the relation of the method to fundamental concepts of the phase space (local, global behavior; interest in approximating specific solutions or studying general qualitative relationships, etc.) The author is not quite blameless in this respect, and the same holds for the discusser (see discussion following reference 2).

It has been frequently stated in the literature (as is also implied by the author) that the phase-plane "method" is useless for systems of higher than second order. There is no serious reason for such pessimism. If the piecewise linear approximation used by both the author and the discusser is agreed to, then in studying third, fourth, and higher order differential equations one can employ all the reasonable approximations used in linear systems. Usually, such approximations can be made to lead to an equivalent second-order problem in the phase plane. This technique has been applied successfully by the discusser[2] to a number of examples of considerably greater complexity than those mentioned in the paper. Therefore the discusser would like to urge strongly that all approximations made in the analysis of nonlinear systems, such as in the present paper, should be carried out either in terms of phase-space concepts or tied to the latter by some means. Only when an analysis can be related to the phase space has a nonlinear problem, namely the study of global behavior, been really solved.

The preceding observations may now be pinpointed as follows. There are two aspects of the analysis of a nonlinear dynamic system: (1) to get an actual time-response curve under stated conditions; (2) to obtain a bird's eye view of the totality of solutions under all admissible conditions; a view distant enough so that it will not be confused by irrelevant detail, yet sharp enough so that it will recognize factors of importance. At the present time, task 1 is becoming less and less essential, since modern methods of machine computing, simulations, etc. have made the computation of specific solutions of differential equations a routine task. Task 2 is still relatively unexplored and not very well understood; yet efficient use of machine computing facilities depends crucially on progress made in this area.

The discusser regrets to conclude that the present paper falls largely under task task 1. Whether it is capable of handling task 2 has not been convincingly demonstrated. At any rate, it seems unlikely that such an extension of the method in the paper can be had without reference to phase space-concepts. On the other hand, application of the ideas of the paper to the study of dynamic systems in the phase space is not only possible but seems quite attractive to the discusser. Many further avenues are open but the present value of Dr. Chen's method appears quite inadequate.

REFERENCES

1. PHASE-PLANE ANALYSIS OF AUTOMATIC CONTROL SYSTEMS WITH NONLINEAR GAIN ELEMENTS,

R. E. Kalman. *AIEE Transactions*, vol. 73, pt. II, 1954, pp. 383–90.

2. ANALYSIS AND DESIGN CONSIDERATIONS OF SECOND AND HIGHER ORDER SATURATING SERVOMECHANISMS, R. E. Kalman. *Ibid.*, vol. 74, pt. II, Nov. 1955, pp. 294–308.

3. INTRODUCTION TO NONLINEAR MECHANICS (book), N. Minorsky, J. W. Edwards. Ann Arbor, Mich., 1947.

4. LECTURES ON DIFFERENTIAL EQUATIONS (book), S. Lefschetz. Princeton University Press, Princeton, N. J., 1946.

Kan Chen: I wish to express my appreciation to Mr. Kalman for his lucid and interesting discussion on my paper. I would like to point out at the outset of my closure that Mr. Kalman's comments are relevant not only to the method proposed in my paper, but also to all approaches used for attacking the problem of designing nonlinear feedback-control systems. Thus his discussion should be of interest to all in the field of nonlinear control.

It is agreed by most people interested in nonlinear servomechanisms, including the authors of some recently published textbooks on advanced servo theory and practice,[1,2] that the two approaches to nonlinear control problems with practical value are the quasi-linearization approach and the phase-space approach. Since the method described in this paper is regarded as a technique pertaining to the quasi-linearization approach, the comments on the method by Mr. Kalman, who has recently made major contributions pertaining to the phase-space approach,[3,4] are particularly welcome. While many points in Mr. Kalman's comments are pertinent, I am unable to agree with Mr. Kalman on his evaluation of the quasi-linearization approach.

From my point of view, the assumption of periodic oscillations in a nonlinear system under the steady state is not *a priori* as Mr. Kalman contends. A system whose trajectory has a stable limit cycle in the phase-space will oscillate with a nonzero finite amplitude. When the trajectories converge to a stable node or focus, the system may be considered as oscillating with a zero amplitude under the steady state. Conversely, as the trajectories diverge from an unstable node or focus towards infinity, then the system may be considered as oscillating with an infinite amplitude under the steady state. Thus the assumption of a periodic oscillation in a nonlinear system under the steady state is not inconsistent from a logical point of view. The assumption used in the describing-function method that this oscillation is sinusoidal is, of course, an approximation which is justified only under certain conditions. This is why the describing-function method may falsely predict the existence of a limit cycle when the transfer-function contour and the describing-function contour intersect at a small angle in the complex S-plane, just as any rough approximation procedure would produce error in the limiting cases. Another point Mr. Kalman seems to have overlooked is that the major practical value of the describing-function method is to guide the designer in reshaping the transfer-function contour in order to achieve stability rather than to predict the amplitude and frequency of the limit cycle. The application of the describing-function method in presenting a picture of the

"global behavior" of a nonlinear system and suggesting a means of modifying the global behavior, so far as stability is concerned, has been demonstrated in a number of papers.[5] It is this practical value of the describing-function method which motivated the development of the method in this paper in an attempt to apply the quasi-linearization techniques to transient study. Further encouragement in developing the method was received as Dr. R. C. Booton successfully extended the idea of quasi-linerization to the analysis of nonlinear systems with random inputs[6] which, incidentally, has never been attacked with the phase-space approach, so far as I know.

Mr. Kalman is correct in pointing out that the method described in the paper is basically a step-by-step procedure. However, the steps of approximation are quite simple after the designer develops a dexterity in estimating transient response of linear systems by the rules of thumb outlined in the Appendix. Furthermore, it should be pointed out that equations 10–15 were written in order to obtain a semi-quantitative solution for the transient problem, which was checked with some published results,[7] thereby confirming the validity of the method. If it were only desirable to obtain a qualitative or global view of the transient problem, the quasi-linearized system in Fig. 15(B) could have been obtained after a simple sketch of $a(t)$ and the pole-zero plots in Fig. 17(A–D) could have been drawn immediately. Once the pole-zero plot of the quasi-linearized system is obtained, the effect of nonlinearity is quite obvious to anyone equipped with a basic knowledge of linear feedback-control theory.

The assumption of step input in the paper was made for the sake of simplicity in illustrating the method. If the input is a ramp, a parabolic, or any other arbitrary function, the method is still completely valid since the actual input may be regarded as a step response. The original system may be modified to take this into account simply by inserting a fictitious linear network preceding the actual command or input point. No special complexity will be introduced by this modification since the method works for an *nth* order system as easily as for a second-order system. This has been demonstrated as part of my thesis work[8] in designing an a-c regulating system, where the disturbance is a combination of step and exponential functions.

The major shortcoming of the proposed method is admitted to be its inadequacy for studying the settling time of nonlinear systems in which the operating point of the nonlinear element oscillates a number of times between two piece-wise linear regions. Fortunately, this state of affairs is usually to be avoided. Even when it exists, the method is still useful in studying the global behavior of the nonlinear system during its early stage of transient, which is characterized by the overshoot and response time of the system. I think there is a possibility of overcoming the afore-mentioned shortcoming of the proposed method by further development of the quasi-linearization techniques, such as testing the nonlinear elements by neither a sinusoid, nor a pulse, but an exponentially damped sinusoid.

Mr. Kalman's suggestion of referring any new method for nonlinear study to the

phase-space concept as the only means of establishing the soundness of the new method is questionable. I agree with Mr. Kalman that the phase-space is a good way, and perhaps the only way, of representing all possible solutions of a differential equation, complete from the early transient to steady state. However, the presentation of a panorama of all possible solutions of a differential equation does not necessarily provide a good clue as to how the form of the given differential equation should be modified in order to yield more desirable practical results. The basic advantage of quasi-linearization approach over the phase-space approach, as I see it, is that while the insertion of a frequency sensitive network in a nonlinear system means the extension of the phase-space to additional dimensions which are difficult to visualize and manage, it means only the addition of a few poles and zeros to the quasi-linearized system on the complex S-plane, or a few break frequencies in the Bode plot, which are always two dimensional. Furthermore, the scale of the independent variable, or time t, is not explicitly displayed in the phase-space. For a second-order system with straight-line segment approximation to the nonlinearity, the time scale may conceivably be calibrated on the standard forms of trajectories. It will be difficult to do the same, however, for an nth order system. By contrast, the time scale is directly connected to the scale of the S-plane and the frequency plot, which are the working media the quasi-linearization techniques.

In spite of my criticisms on the phase-space approach, I should leave no impression that I have ignored the mathematical rigor which underlies the phase-space approach,[9] the perspective of the global behavior of a given system presented by the phase-space approach[3,9] and its amenability to optimum relay servo design.[10] On the contrary, I believe the present efforts of mathematicians and engineers in the phase-space approach[4,11,12] will yield fruitful results in the field of nonlinear control. Furthermore, I believe that cross reference of the quasi-linearization approach and the phase-space approach will bring about a better understanding of the nonlinear problem and perhaps a basically new engineering method of attacking the problem. However, I do not agree with Mr. Kalman that the quasi-linearization approach must be referred to the phase-space approach before it can have any value for nonlinear servo design.

REFERENCES

1. AUTOMATIC FEEDBACK CONTROL SYSTEM SYNTHESIS (book), J. G. Truxal. McGraw-Hill Book Company, Inc., New York, N. Y., 1952.

2. SERVOMECHANISM AND REGULATING SYSTEM DESIGN (book, vol. 2), H. Chestnut, R. W. Mayer. John Wiley and Sons, Inc., New York, N. Y., 1955.

3. See reference 1 of the discussion by Rudolf E. Kalman.

4. See reference 2 of the discussion by Rudolf E. Kalman.

5. See references 1–4 of the paper.

6. See reference 5 of the paper.

7. See reference 8 of the paper.

8. See reference 6 of the paper.

9. See reference 3 of the discussion by Rudolf E. Kalman.

10. PHASE-PLANE APPROACH TO THE DESIGN OF SATURATING SERVOMECHANISMS, A. M. Hopkin. AIEE Transactions, vol. 70, 1951, pp. 631–39.

11. A METHOD FOR SOLVING THIRD AND HIGHER ORDER NONLINEAR DIFFERENTIAL EQUATIONS, Y. H. Ku. Journal, The Franklin Institute, Philadelphia, Pa., vol. 256, no. 3, Sept. 1953, pp. 229–44.

12. ACCELERATION METHOD PLANE FOR ANALYSIS OF A CIRCUIT WITH NONLINEAR INDUCTANCE AND NONLINEAR CAPACITANCE, Y. H. Ku. AIEE Transactions, vol. 73, pt. I, 1954, pp. 619–26.

13. BACKLASH IN VELOCITY-LAG SERVOMECHANISM, N. B. Nichols. Ibid., vol. 72, pt. II, 1953, pp. 631–39.

Dynamics of D-C Machine Systems

M. RIAZ
ASSOCIATE MEMBER AIEE

Synopsis: A generalized block diagram representation of the signal transfer properties of d-c machine systems with various kinds of feedback is developed in terms of gain factors, time constants, and feedback ratios, including the mutual coupling effects between windings and the nonlinear effects produced by saturation. The resulting formulation of the dynamic properties is simple enough to permit a rapid evaluation of the action of feedback in d-c machine systems and therefore to help in determining which feedback combinations will secure a desired system performance.

THE science of control engineering has made rapid progress in recent years. As the need for more accurate and reliable control grew, demands arose for machines that would possess characteristics suitable for use in control systems.

One aspect of the development in d-c machines which normally constitute the power-amplifying stages of a control system emphasizes the design of machines capable of responding very fast to signals of small magnitude. These dynamoelectric amplifiers, often known by trade names such as Amplidyne, Rototrol, Regulex, VSA, Magnavolt, etc., have been successfully employed in many control systems. Another aspect, of a more general nature, features the introduction of feedback of various forms both inside and outside the machine to control its characteristics, and leads to the development of the metadyne as a generalized commutator machine.

Although many of these machines have been known for a long time, it is only recently that their dynamic characteristics were closely studied.[1,2] For the type of problems met in control systems, the dynamic properties are usually expressed in terms of transfer functions (input-output operational relationships) or frequency response characteristics, according to the well-known techniques of feedback control systems. However, from an analytical viewpoint, the handling of internal feedback in machines is considerably involved, mainly because of the mutual coupling effects between field windings. The resulting expressions often become so complex that it is difficult to compare the performance of different systems or to predict the influence of parameter or feedback variations upon the over-all system behavior. Another difficulty arises in those problems where saturation nonlinearities must be included in the analysis as, for example, in a voltage regulatory problem for which faults, switching operations, and load changes would make a linear treatment invalid.

It is the purpose of this paper to present a general formulation of the dynamic properties of d-c machine systems which takes into account both mutual inductive effects and saturation nonlinearities. This formulation is characterized by the simplicity that is achieved by making judicious assumptions which are closely verified in practice and also by expressing the results in block diagram form. The difficulty of accounting for the mutual effects vanishes if the coupling between windings is assumed perfect. The nonlinear effects can be handled simply if the saturation of the magnetic circuit alone is present, while the remainder of the machine system may still be assumed linear. This permits the isolation of the nonlinearity and its inclusion in an otherwise completely linear representation of the system.
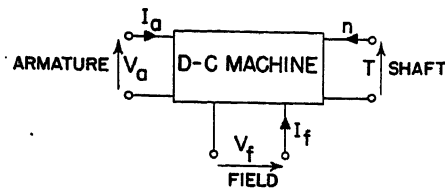
Fig. 1. D-c machine as a power modulator

Nomenclature

V_k = signal voltage to field circuit labeled k
V_0 = output voltage at base speed n_0
i_k = current in winding k
$Z_k(p)$ = equivalent internal impedance of field coupling network and source looking back from winding k and including the resistance R_k of that winding
$O_k(p)$ = open-circuit transfer ratio of coupling network connected to winding k
p = time differential operator d/dt
n = speed in revolutions per second
ϕ = air-gap flux per pole
σ = coefficient of dispersion
m = number of poles
N_k = number of turns per pole of winding k
z = total number of armature conductors
a = number of parallel paths in armature
P = permeance of a single air gap
L_k = unsaturated inductance of winding k
R_k = resistance of winding k
K_k = linear speed-voltage coefficient of winding k at base speed n_0
S = saturation function
$T_k = L_k/R_k$ = time constant of winding k
$\alpha_k = K_k/R_k$ = zero-frequency voltage gain relative to field winding k

The q-Field D-C Generator

For control purposes, a d-c machine is best viewed as a power modulator, a device in which power flow from a source (reservoir) to a load is modulated or controlled under the dictates of a signal (command). It is characterized by at least three sets of terminal pairs, shown in Fig. 1, which define the field, the armature, and the shaft. Since the process of electromechanical energy conversion is reversible, the flow of power between armature and shaft is bilateral. Modulation of this power flow is obtained by acting on the third terminal pair, the field. How-
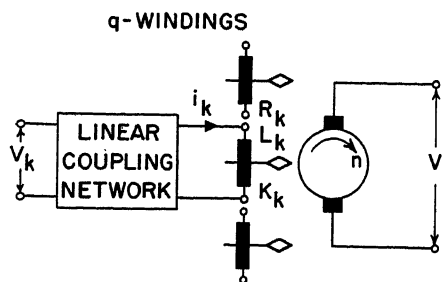
Fig. 2. General configuration of q-field d-c generator

ever, in order to control more closely the characteristics, feedback is introduced both inside and outside the machine so that, in general, this power modulator normally appears with more than three terminal pairs: several field windings as well as several armature circuits may be provided in a single machine. The multi-field d-c machine with a single armature terminal pair, as shown schematically in Fig. 2, will first be examined; it will serve as a basic generalized configuration from which any particular arrangement of a d-c machine may be derived by suitable interconnections and feedback combinations. Multiarmature machines, like metadynes, are briefly treated in the Appendix.

The signal transfer properties of the q-field d-c generator will now be obtained on the basis of the following assumptions:

1. The q-field windings are perfectly coupled. This assumption is justified in practice since all field windings are normally placed on the same pole structure and link the same total field flux. It will further be assumed that a fixed proportion of this flux links the armature; to the machine is therefore assigned a constant coefficient of dispersion σ, defined as the ratio of total field flux to air-gap flux.

2. The brushes are located in the neutral zones. Commutation is assumed to be linear, and the effects of the coils undergoing commutation are ignored.

3. The magnetic properties of the direct (field) and quadrature (armature) axes are taken to be mutually independent. The armature internal impedance is considered to be lumped with the load that the generator is supplying.

Furthermore, the following information is given:

1. The magnetization curve representing the relation between generated voltage at a constant base speed n_0 and field ampere-turns is known.

2. Any given coupling network interposed between a signal voltage and a field winding behaves linearly, so that it can be represented by a Thevenin equivalent, as shown in Fig. 3.

The following differential equations can be written, applicable to every field circuit

$$V_k O_k(p) = Z_k(p)i_k + m\sigma N_k p\phi$$
$$(k = 1, 2, 3 \ldots q) \quad (1)$$

The straightforwardness of the ensuing derivations results from the equations written on essentially a nodal basis: in fact, in terms of one node variable, the air-gap flux. However, since the magnetization curve is plotted in terms of voltage, it is convenient to write the differential equations in terms of voltage instead of flux by letting
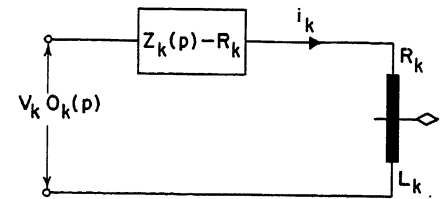
Fig. 3. Thevenin circuit equivalent of winding k

$$V_0 = cn_0\phi \quad (2)$$

where c = proportionality constant = mz/a.

Hence

$$V_k O_k(p) = Z_k(p)i_k + \frac{m\sigma N_k}{cn_0} p V_0 \quad (3)$$

from which

$$N_k i_k = V_k N_k \frac{O_k(p)}{Z_k(p)} - \frac{m\sigma N_k^2}{cn_0} \frac{1}{Z_k(p)} p V_0 \quad (4)$$

After summing the ampere-turns for all q windings, with due regard to polarities, the resultant ampere-turns per pole are

$$\sum_{k=1}^{q} N_k i_k = \sum_{k=1}^{q} V_k N_k \frac{O_k(p)}{Z_k(p)} - \frac{p V_0}{cn_0} \sum_{k=1}^{q} \frac{m\sigma N_k^2}{Z_k(p)} \quad (5)$$

The slope of the linear portion of the magnetization curve (the no-load characteristic) is $cn_0 P$, where P is the permeance of a single air gap.

After recognizing that in the linear region $cn_0 P N_k$ is the linear speed-voltage coefficient K_k of the field winding k in volts per ampere at base speed n_0, and that $m\sigma P N_k^2$ is the unsaturated inductance L_k of winding k, then multiplication of equation 5 by $cn_0 P$ yields

$$\sum_{k=1}^{q} K_k i_k = \sum_{k=1}^{q} V_k K_k \frac{O_k(p)}{Z_k(p)} - p V_0 \sum_{k=1}^{q} \frac{L_k}{Z_k(p)} \quad (6)$$
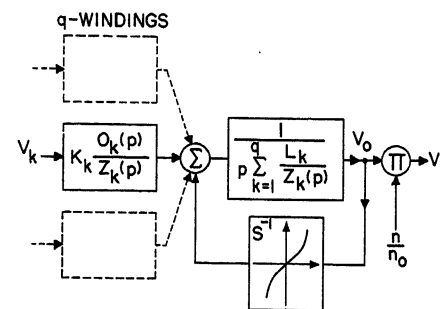
Fig. 4. Block diagram representation of q-field d-c generator

Riaz—Dynamics of D-C Machine Systems
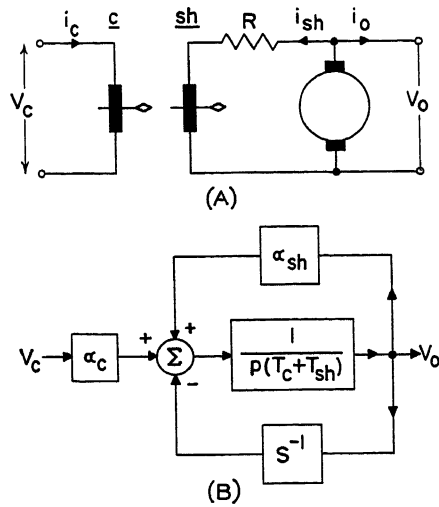
Fig. 5. Shunt feedback in d-c generator

A—Machine configuration
B—Block diagram representation

Similarly, from equation 2, we have for the linear region

$$V_0 = cn_0 P \sum_{k=1}^{q} N_k i_k = \sum_{k=1}^{q} K_k i_k \qquad (7)$$

and, in general, when saturation is present

$$V_0 = S(\sum K_k i_k)$$

or $\sum K_k i_k = S^{-1}(V_0)$ (8)

where S represents a saturation function with unity initial slope which corresponds to the linear portion of the magnetization curve, and S^{-1} denotes the inverse saturation function. If the d-c machine is running at a speed n different from n_0, the output voltage becomes

$$V = V_0 \frac{n}{n_0} \qquad (9)$$

These various relations are best expressed in terms of block diagrams because the representations provide a more effective way for observing the interactions of diverse parts of a machine than the original differential equations or transfer functions. Equations 6, 8, and 9 lead directly to two alternate block diagram representations of the q-field d-c machine. One of these is drawn in a simple form as Fig. 4. The other representation is essentially the inverse, in which a saturation function S is placed in the forward path and a differentiator in the feedback path. Despite the complete mathematical equivalence of such a representation to that shown in Fig. 4, it is not suited for analogue computer applications because of the practical difficulty of performing the differentiation process.

As indicated in Fig. 4, one of the two inputs to the multiplication box placed at the output is a per-unit speed which, in general, may vary in time, depending on the dynamics of the drive. If, as is often the case, the speed can be assumed constant, the multiplication box is then simply replaced by a gain factor.

In many control problems, a system may be linearized around an operating point, provided only small perturbations are considered. The question of stability, in particular, may be handled by considering incremental quantities only. The relations between these incremental changes can be expressed in a linearized block diagram obtained from Fig. 4 by replacing the feedback nonlinear block S^{-1} by an incremental gain factor equal to the slope of the inverse saturation function at the operating point. In particular, if the operation is restricted to the linear region of the magnetization curve, the feedback in Fig. 4 is unity. In this case, the over-all multiple-input to output transfer function is

$$V_0 = \frac{\displaystyle\sum_{k=1}^{q} K_k \frac{O_k(p)}{Z_k(p)} V_k}{1 + p \displaystyle\sum_{k=1}^{q} \frac{L_k}{Z_k(p)}} \qquad (10)$$

Illustrative Examples

Consider first the simple case of the q-field d-c generator in which the signal voltages are all directly impressed onto the field windings. The linear transfer function given as equation 10 becomes in this instance

$$V_0 = \frac{\displaystyle\sum_{k=1}^{q} (K_k/R_k) V_k}{1 + p \displaystyle\sum_{k=1}^{q} L_k/R_k} = \frac{\displaystyle\sum_{k=1}^{q} \alpha_k V_k}{1 + p \displaystyle\sum_{k=1}^{q} T_k} \qquad (11)$$

where $\alpha_k = K_k/R_k$, $T_k = L_k/R_k$, and V_k are the voltage gain (at zero frequency or d-c condition), time constant, and signal voltage of the kth field circuit respectively. Equation 11 indicates that the q-field d-c machine behaves dynamically as a single time-lag system in which the equivalent time constant is equal to the sum of the individual time constants for each field winding.[3]

If saturation is present, the block diagram of Fig. 4 shows that the resulting effect is an equal diminution of both gain and time constant. At any instant of time corresponding to a certain operating point on the saturation curve, the gain and time constant are reduced by a

factor equal to the slope of the saturation function curve at the instantaneous operating point.

The effects of eddy currents in the field structure of the machine may be conveniently included in the analysis if represented by an equivalent short-circuited field winding having appropriate resistance and inductance values.

In the Appendix, the generalized block diagram is employed to derive the dynamic characteristics of several typical machine configurations. Three basic sets of configurations are briefly examined:

1. Scalar (both shunt and series) feedbacks in a single-stage machine.
2. Cascade operation of d-c machines (and multistage machines).
3. Feedback through frequency-sensitive networks.

The object of presenting these various examples is not to develop a more or less complete analysis of any one particular configuration but rather to indicate how various techniques of feedback control systems can be fruitfully adapted to the study of d-c machine systems. A Rototrol, or a Regulex, or an Amplidyne system (to mention only well-known dynamoelectric amplifiers) is here viewed as being in each case a definite combination of feedback yielding specific characteristics expressible in such forms as block diagrams, transfer or describing functions, frequency responses, or root loci.

Conclusion

A unified approach to d-c machine dynamics has been presented in terms of feedback concepts leading to a generalized block diagram representation of these systems. The basic assumptions are few and, in practice, nearly always justifiable. Such important effects as saturation, variable speed, and loading or mutual coupling are included in the formulation of the system dynamics. This complete yet relatively simple formulation points out how feedback may be manipulated in d-c machines to determine from a comparative standpoint the behavior of various types, or to secure specified system performances. Several feedback arrangements have been selected in this paper to demonstrate some known machine configurations; many more exist or may be devised. The choice of the optimum configuration and feedback to be used for any specific application remains the challenging problem (the synthesis problem) facing the designer of d-c machinery for control systems.
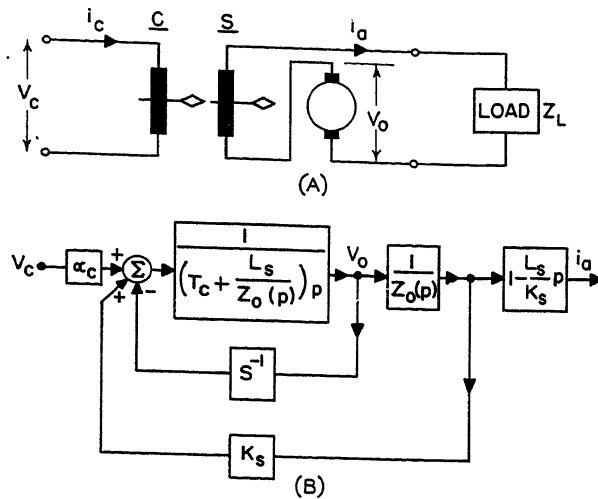
Fig. 6 (left). Series feedback

A—Machine configuration
B—Block diagram representation
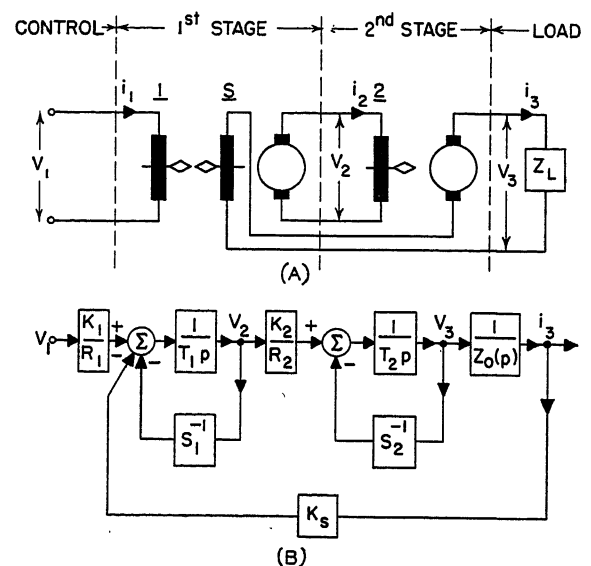
Fig. 7 (right). Sequence of two machines with negative series feedback

A—Machine configuration
B—Block diagram representation

Appendix. Dynamics of Some Typical D-C Machine Configurations

The multiplicity of configurations and hence the flexibility of the resulting characteristics evolve from the various ways in which feedback may be utilized. Several basic kinds of feedback may be distinguished in a machine system: external feedback and internal (with the provision of extra windings inside the machine); scalar (direct) feedback which comprises series (current) and shunt (voltage) types, and then feedback through frequency-sensitive elements; positive (regenerative or cumulative), and negative (degenerative or suppressive) feedback. These various kinds of feedback may in turn be classified in two essential groups: operational and modifying. Operational feedback improves the performance of a machine but does not substantially modify its characteristics. Examples are the passing of armature current through interpole windings to improve commutation, and through pole-face (compensating) windings to neutralize the cross-magnetizing effects of armature reaction, particularly in d-c machines subjected to violent fluctuations in load. The reduction in armature inductance resulting from the addition of a compensating winding is an incidental effect that can be readily taken into account in the analysis. Operational feedback will not be considered here but some types of modifying feedback, by means of which the steady-state and transient characteristics of d-c machine systems can be controlled and adapted to meet given specifications, will be examined.

Shunt Feedback

A d-c machine arrangement with positive shunt (voltage) feedback is illustrated in Fig. 5(A), and its block diagram for constant speed is drawn as Fig. 5(B). This machine is equipped with two field windings: one, denoted by c, is separately excited and is called the control winding; the other, labeled sh, is the shunt winding connected across the machine output terminals. The generator can operate with either

field acting alone or with both excitations present.

As is well known, the operation of the generator with shunt excitation only is inherently dependent upon the nonlinearity of the magnetization curve. This condition can be deduced directly from a consideration of the block diagram of Fig. 5(B). (Winding c in this case may be short-circuited and assumed to simulate the eddy-current effects.) Another required condition is that the gain $\alpha_{sh} = K_{sh}/(R_{sh}+R)$ must be larger than unity. The steady-state characteristics are obtained from the block diagram representation by setting $p=0$. Some residual or initial voltage of suitable polarity must exist in order to build up the excitation of the machine. The output voltage is controlled by adjusting the resistance R in the shunt circuit, i.e., varying α_{sh}.

Consider next the operation of the generator when a signal voltage is fed in its control field c. The action of the control winding is to add or subtract a certain number of ampere-turns from those introduced by the main shunt field. In effect, this produces a stabilizing action which extends the operation of the machine to its linear region and permits the adjustment of the gain α_{sh} to values less than unity. For these conditions, inspection of Fig. 5(B) yields the over-all transfer function

$$\frac{V_0}{V_c} = \frac{\alpha_c}{1-\alpha_{sh}} \frac{1}{\left[\dfrac{T_c+T_{sh}}{1-\alpha_{sh}}p+1\right]} \quad (12)$$

This relation shows that the increase in gain accrued from positive shunt feedback is accompanied by an identical increase in equivalent time constant. It also indicates the delayed response for a change in control voltage that is caused by the mutual inductive coupling between control and main shunt fields. The presence of T_{sh} in addition to T_c may thus be interpreted as due to the damping action taking place in the shunt winding.

The loop gain, which is equal to the gain associated with the shunt winding $\alpha_{sh} = K_{sh}/(R_{sh}+R)$, is less than unity if self-excitation of the machine is to be avoided. In some cases, however, this loop gain is purposely set equal to unity by adjustment of the resistance R in the shunt field

circuit, a technique known as tuning. When $\alpha_{sh}=1$, the transfer function becomes

$$\frac{V_0}{V_c} = \frac{\alpha_c}{T_c+T_{sh}} \frac{1}{p} \quad (13)$$

which indicates that the generator acts as a perfect integrator. Mathematically, this means an infinite steady-state gain. In practice, the gain is not infinite but can be quite large, though very sensitive to slight changes in speed and resistance, which affect the tuning condition $K_{sh} = R_{sh}+R$. Machines operating on this principle have been used extensively and are often known by trade names such as Regulex (Allis Chalmers Manufacturing Company), VSA (Reliance Electric and Engineering Company), and Magnavolt (English Electric Company)[2] and HTD (Hitachi, Ltd.; Japan).
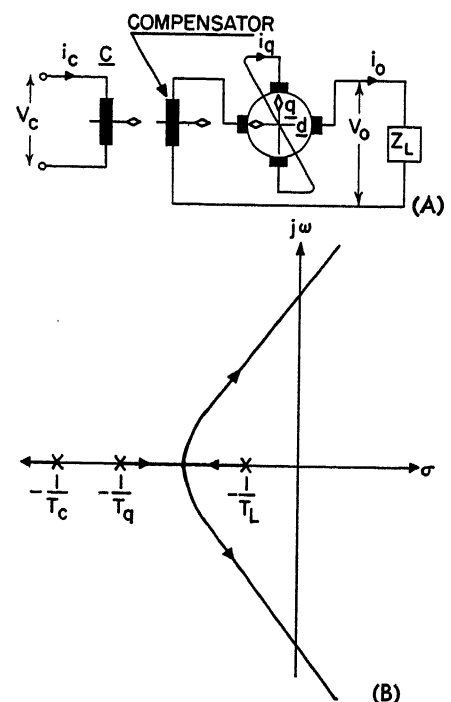
Fig. 8. Metadyne system

A—Machine configuration
B—Root locus

Series Feedback

When series feedback is employed, i.e., armature current passed in a feedback winding, the load circuit must be included in the dynamic analysis of the d-c machine. A general configuration of series positive feedback is illustrated in Fig. 6, together with the equivalent block diagram representation. In the diagram, $Z_0(p)$ refers to the whole armature circuit impedance excepting the inductance L_s of the series winding. From a dynamic standpoint, three different types of loads can be distinguished:

1. The resistive load.
2. The inductive load encountered, for example, when the d-c generator furnishes excitation to the field of another machine.
3. The capacitive load, as in the Ward-Leonard control of a d-c motor.

The loading effect created by the feedback winding upon the control circuit appears in the block diagram as the term $L_s/Z_0(p)$. This term can be ignored if the series field inductance is small compared with the load impedance. The effect of the voltage induced in the load circuit is indicated by the block $[1-(L_s/K_s)p]$ which results from applying equation 4 to the present machine configuration

$$i_a = \left[1 - \frac{L_s}{K_s}p\right]\frac{V_0}{Z_0(p)}$$

To confine attention to the linear operation of the d-c machine in the presence of an inductive load, a loop gain may be defined as $\alpha_s = K_s/R_0$, which is normally less than unity. The over-all zero-frequency gain is then $\alpha_c/(1-\alpha_s)$. If α_s is made equal to unity by tuning the series load circuit, the gain becomes infinite, as evidenced by the integration in the transfer function

$$\frac{i_a}{V_c} = \frac{\alpha_c}{R_0(T_c+T_L)}\frac{\left(1-\frac{L_s}{K_s}p\right)}{p\left(\frac{T_cT_L}{T_c+T_L}p+1\right)} \quad (14)$$

A machine constructed on this principle is sometimes known as a Rototrol[2] (Westinghouse Electric Corporation).

Both shunt and series direct feedback discussed so far have been of the positive or regenerative type. If, instead, negative feedback is resorted to in a d-c machine, the effects are just the opposite and may be taken care of in the block diagrams of Figs. 5 and 6 by making appropriate changes in sign. Negative feedback increases the speed of response of the machine and often has a stabilizing tendency in a control system, since it effectively reduces the time lags. This improvement is, however, obtained at the expense of decreased amplification.

Cascade Arrangements of D-C Machines

The power amplification of a d-c machine can be increased either by having a larger field structure or by introducing positive feedback, but this means an unavoidable increase in time constant. However, in combining several machines in cascade, a large over-all amplification can be achieved at the expense of the comparatively smaller time lags associated with each machine of the sequence. Feedback arrangements of even the simplest types are innumerable when applied to multistage machines. Within the limits of the basic assumptions, the block diagram approach presented in this paper can be extended to cover multistage machine systems.

As an example, Fig. 7 illustrates the case of a 2-stage machine amplifier with series negative feedback across both stages. The coupling effects between control and series windings in such an arrangement, producing a large amplification, are usually of such small magnitudes that the simple block diagram shown in Fig. 7(B) is a valid representation of the system dynamics. This system can also be realized in a single machine of the cross-field construction. Often known as a metadyne,[4] it essentially combines in one armature-excited machine two stages of amplification with negative current feedback from the output. A basic arrangement of a metadyne system is shown in Fig. 8(A), for which, with certain qualifications, the block diagram of Fig. 7(B) is still applicable. First, it must be recognized that the original assumption of perfect coupling between field windings may be not as well justified in a construction involving armature and field couplings. Second, caution should be exercised with regard to saturation. In the representation of a metadyne such as Fig. 7(B), it is implied that saturation in any one axis does not interfere with saturation in the other axis. Although this assumption introduces a great deal of simplicity in handling this nonlinear effect, it may cause, in some specific cases, appreciable errors in dynamic system prediction. The root locus of the metadyne system shown in Fig. 8(B) leads to a quick understanding of the effect output compensation upon system dynamic performance. With zero loop gain, both open and closed loop poles coincide, resulting in an overdamped transient behavior. This condition is realized with perfect compensation (i.e., no feedback when $K_s=0$) and corresponds to the amplidyne configuration.[2] As the amount of compensation is reduced, or in equivalent terms, as negative feedback is increased, the loop gain increases so that the closed-loop poles follow the root loci in the arrow directions, eventually yielding an underdamped behavior and an increase in speed of response. For a sufficiently large loop gain, the system may even become unstable.

Feedback Through Frequency-Sensitive Networks

Feedback through frequency-sensitive elements constitutes another important class of feedback which is utilized to improve the dynamic characteristics of machine systems. In particular, it offers a means of incorporating stabilizing features to these systems. In contrast with electronic systems, the cascade type of stabilization does not normally provide a practical solution to the stabilization problem at the power levels encountered in machine components. Transient feedback from intermediate stages with the use of transformer or capacitor arrangements is a well-known damping technique. A somewhat less conventional scheme involves internal feedback derived from an auxiliary winding placed on the poles of a d-c machine. The general block diagram of Fig. 4 can effectively be used to describe the action of feedback through frequency-sensitive networks.[5]

References

1. Direct-Current Machines for Control Systems (book), A. Tustin. The MacMillan Company, New York, N. Y., 1952.

2. Bibliography on Feedback Control, AIEE Committee Report. *AIEE Transactions*, vol. 72, pt. II, 1953 (Jan. 1954 section), list for "Dynamo Amplifiers," pp. 458-59.

3. Transient Operation of D-C Generators, D. Journeaux. *Allis-Chalmers Electrical Review*, Milwaukee, Wis., vol. 15, no. 3, 1950, p. 27.

4. Transient Analysis of the Metadyne Generator, M. Riaz. *AIEE Transactions*, vol. 72, pt. III, Feb. 1953, pp. 52-62.

5. Alteration of the Dynamic Response of Magnetic Amplifiers by Feedback, R. O. Decker. *Ibid.*, vol. 73, pt. I, 1954 (Jan. 1955 section), pp. 658-64.

Discussion

Robert M. Saunders (University of California, Berkeley, Calif.): As one who has coped with some of the problems which Dr. Riaz has so adroitly solved, I can appreciate the finesse and utter simplicity of his Fig. 4. The reader should not overlook the universality of this representation.

In the Appendix the author calls attention to the assumption that saturation in any one axis does not interfere with saturation in the other axis and to the the attendant possibility of error if it should. Would this not be a general requirement of all d-c machines whether of the simple single-field 2-brush-per-pole type or of the more complex metadyne type? To phrase the question somewhat differently, does the representation account for the demagnetizing effects in the direct axis and consequent coupling between direct and quadrature axes under saturation conditions? In control system applications such cross-axis influences are not very important, but they might be if an extension to generator fault current analysis were to be made with the use of the representations included here.

In Fig. 1 the author suggests that the same type of thinking applies to d-c motors although the examples in the paper seem to be pointed at generators. I hope he will follow this paper up with one on representations for motors. I should like to emphasize that I consider this paper to be one of the most significant contributions to the d-c machinery literature made in recent years.

M. Riaz: Prof. Saunders' comments are sincerely appreciated. As pointed out in the discussion the situation created by saturation in coupling two or more flux systems occurs in all commutator machines. The paper does not account for it, and the simplicity of Fig. 4 is partly the result of assumption 3, which neglects the interfering action of fluxes in space quadrature.

While this assumption is reasonably warranted in control problems, it is not as well justified in such power problems as the prediction of the short-circuit characteristics of d-c motors and generators. Reference 1 of the paper gives some techniques for including the effects of the cross-axis flux interferences produced by saturation upon the steady-state characteristics of

conventional and cross-field d-c machines. Such techniques may also be used to develop block diagram representations of the machines which account for the effects of the flux interferences due to saturation. These representations are more complicated than those given in the paper since they involve saturation nonlinearities that are functions of two or more variables.

Surge Protection on Industrial Systems

C. L. WAGNER
ASSOCIATE MEMBER AIEE

THE rapid expansion of industrial plant distribution systems in recent years has caused many new problems for the plant design engineer. It has also accentuated certain problems that in the past caused little concern. Surge protection for the plant system is one of these latter problems.

In the past, the philosophy behind surge protection was primarily to prevent damage to the electric equipment itself and thereby keep the repair costs for this equipment to a minimum. With this philosophy, not too much emphasis was placed on surge protection by the over-all plant designers. In recent years, however, continuity of service has become a paramount concern. Now the breakdown of a piece of equipment means not only repair or replacement expense but, more importantly, loss of production for that portion of the plant affected by the breakdown. In many cases, primarily in the chemical industry, it could mean the complete shutdown of the plant because of the interruption of a critical processing operation. Therefore, more and more emphasis is being placed on surge protection by the plant designers.

Adequate surge protection consists not only of protection against lighting surges but also against transient voltages caused by switching surges or arcing ground faults. In many cases, damage resulting from these transient voltages is far more extensive than that caused by lightning surges. Proper system design, therefore, requires that protection be afforded against both types of surges.

The science of surge protection is one of

prevention plus control. The first step to be taken by the plant designer is to attempt to prevent the occurrence of surges on his system. This might take the form of shielding of his lines and stations to minimize the probabilities of lightning strokes to his system. It might also take the form of proper grounding to prevent the establishment of excessive transient voltages. The second step is to provide a means of controlling the surges that slipped by the preventive schemes. This may take the form of lightning arresters, surge capacitors, or any other equipment necessary to prevent damage to the apparatus on the system. Both steps must be taken to provide the maximum in protection for the plant system. Some of the prevention and control methods available for the protection of plant electric equipment against switching as well as lightning surges are presented in this paper.

Switching Surges

When electric systems are subjected to sudden circuit disturbances, transient voltages or surges are likely to be produced. Circuit disturbances that produce the highest transient voltage involve arc paths. The arc path may be in a circuit breaker or in a line-to-ground fault on the system. In either case, high transient overvoltages are possible, providing the system constants and the arc characteristics satisfy certain conditions. An explanation of switching surges and arcing ground phenomena is given in reference 1.

Protection of the industrial plant system against damage caused by switching surges and other transient voltages is in many cases more critical than protection against lightning surges. When the control step of the prevention-plus-control program of surge protection is attempted, certain difficulties immediately arise. In

the first place, the switching surge strength of electric apparatus has not been definitely established. The frequencies involved in these switching surges vary between 1,000 and 2,000 cycles. Thus, the apparatus insulation strength for these surges probably lies somewhere between the 60-cycle and the impulse-insulation levels. At present, however, this value is not known exactly. The same difficulty arises with the lightning arresters. There are not sufficient data on arresters at this switching surge frequency to determine what protective characteristics they possess. A considerable amount of work is being done by the industry at the present time in an attempt to establish these values.

Since the control portion of the problem is so vague, what can be done in the way of prevention? This portion of the problem is more clean cut. Computer studies as well as experience on actual systems have shown that, by proper system grounding, transient overvoltages caused by arcing grounds or switching conditions can be limited to low enough values so that apparatus insulation will not be endangered.[2,9]

REACTANCE GROUNDING

Fig. 1 shows the effect of system grounding on the magnitude of the transient overvoltages. When the ratio of the system zero-sequence reactance to the positive-sequence reactance X_0/X_1 is less than 10 (the line-to-ground fault current being greater than 25 per cent of the 3-phase fault current), the transient overvoltages are limited to 250 per cent of the normal line-to-neutral crest voltage or less. If X_0/X_1 is greater than 10, the overvoltages may exceed 500 per cent of the normal line-to-neutral crest voltage. This would indicate that, to limit the transient overvoltages, the X_0/X_1 ratio of the system should be kept less than 10. Actually, the results plotted in Fig. 1 were obtained by extinguishing the arc at a current zero, i.e., without forcing current zero. If the fault is in a confined space such as under oil, in cable, or apparatus insulation, the expulsion action of the de-ionizing agents present would tend to

PERCENT OF SURGES EXCEEDING ORDINATES

blow the arc out prior to its normal current zero. This forcing current zero tends to increase the transient overvoltages. Other computer studies were made simulating this condition and it was found that X_0/X_1 ratios as low as 3 or 4 (line-to-ground fault current being greater than 50 to 60 per cent of the 3-phase value) were necessary to limit the transient overvoltages to 250 per cent or less.[1,2]

The computer data plotted in Fig. 1 show the magnitude of the overvoltages to be as high as 500 per cent for certain grounding conditions. Overvoltages even higher than this have been known to occur on actual systems. Fig. 2 shows a summary of the results of two klydonograph investigations on switching surges. This figure shows that voltages as high as 600 per cent of normal line-to-neutral crest voltage have been recorded.[3,4]

Fig. 1 and the foregoing showed that, if reactors were used to ground the system, a value of reactance must be chosen such that the line-to-ground fault current must be not less than 50 per cent of the 3-phase fault current. This magnitude of fault current is liable to cause quite extensive fault damage either in the form of mechanical damage to the equipment or in the form of iron burning in the machines. It might be felt desirable to reduce the magnitude of this current to a more reasonable value. If such is the case, resistance grounding should be used.

RESISTANCE GROUNDING

There are two types of resistance grounding, namely, low and high resistance. The type of resistance grounding used depends on the relaying requirements of the system. If it is desired to use selective ground relaying on the

system, then low resistance grounding must be employed. As shown in Fig. 3, this resistor could be installed in the neutral of a generator or transformer or, if no such neutral point exists, it could be installed in the neutral of a zig-zag or wye-delta grounding transformer. To insure positive ground relay operation, the ohmic value of the resistor should be such that the line-to-ground fault current will be at least equal to the primary current rating of the largest current transformer on the system. Also, to limit the

magnitude of the transient overvoltages, the ratio of the zero-sequence resistance to the zero-sequence reactance R_0/X_0 should always be equal to or greater than 2.0 and X_0/X_1 must be less than 20.

Some system operators do not have ground relaying and do not want to have an outage or trip a feeder when there is just one ground fault on the system. For this reason they operate their system ungrounded. It is on this type of system that high transient overvoltages are most likely to occur. The presence of these overvoltages can also nullify the advantages of ungrounded operation. For example, on an ungrounded system a piece of apparatus might fail and cause an arcing fault to ground. This arcing fault may produce high enough transient overvoltages in other portions of the system for several other pieces of equipment in the plant to fail. These other failures
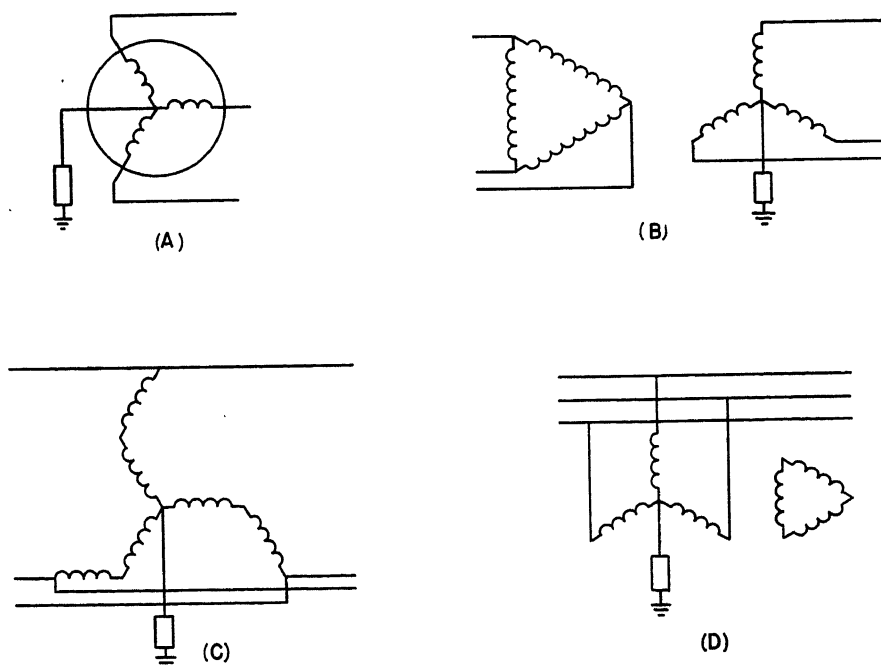
Fig. 3. Low-resistance neutral grounding schemes

A—Generator neutral resistor
B—Power transformer neutral resistor
C—Zig-zag grounding transformer neutral resistor
D—Wye-delta grounding transformer neutral resistor
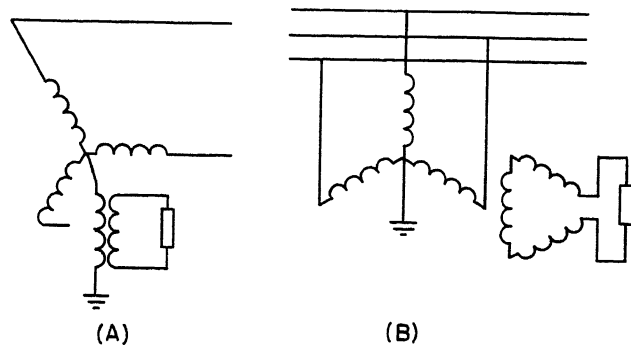
Fig. 4. High-resistance grounding schemes

A—Neutral point available, distribution transformer with secondary resistor

B—No neutral point available, 3-phase distribution transformer with secondary resistor

$$M = \frac{\sqrt{LC}}{2RC} \qquad T = 2\pi\sqrt{LC}$$

Fig. 6. Solutions for determining the voltages from machine terminals to ground produced by electromagnetic transfer through transformer

A—Actual circuit
B—Equivalent circuit
E_a—High-voltage arrester discharge voltage
V_m—Voltage at machine terminals
L—Transformer leakage inductance
C—Capacitance on machine side of transformer
R—Effective surge impedance of machine
N—Ratio of the machine to line voltages

may involve different phases which would mean that a line-to-line fault would result and the involved feeders would still have to be tripped immediately. Thus the main advantage of ungrounded operation would be eliminated.

To eliminate the transient-overvoltage problem and still have the advantages of the ungrounded operation, high-resistance type of grounding is recommended. In this case, a value of resistance is selected such that, under fault conditions, the kilowatt loss in the resistor is equal to the capacitive charging kilovolt-amperes of the system. With this value of resistance, the transient overvoltages are limited to 250 per cent of the normal line-to-neutral crest voltage. With this value, also, the fault current is only 50 per cent greater than when the system was operated ungrounded. Thus, the over-all operation of the system would be no different than when ungrounded, the only exception being that the trouble with transient overvoltages would be eliminated.

Fig. 4 shows how this resistor could be installed on the system. If a neutral

Fig. 5. Simplified schematic of a power transformer

C_H—Lumped capacitance of high-voltage winding to ground
C_L—Lumped capacitance of low-voltage winding to ground
C_M—Lumped capacitance between windings

point is available, it should be installed in the secondary circuit of a distribution transformer which, in turn, is connected from the neutral point to ground. This transformer should have a voltage rating equal to the line-to-line voltage of the system. If a neutral point is not available, the resistor should be installed in the secondary circuit of a wye-broken delta distribution transformer. Again the line-to-neutral rating of the distribution transformer should be equal to the system line-to-line voltage. In both schemes a voltage relay could be connected across the resistor to sound an alarm when there is a fault on the system.

This latter type of grounding is gaining wide acceptance in the industry. One company, who has a number of plants throughout the country, until recently operated all their systems ungrounded. A very high rate of motor failures were reported by their plants, and lightning was suspected as being the cause of the damage. On examining the trouble reports, it was found that in most cases multiple failures occurred. In many cases the simultaneous failures were located quite some distance from each other. This

led to the suspicion that switching surges or arcing grounds were the cause of trouble and not lightning. The high-resistance grounding scheme was applied to all their plant systems and, at the last report, the motor-failure rate had dropped to a small fraction of what it was before the systems were grounded.

It should be kept in mind that, whenever resistance grounding is used, the neutral is fully displaced during fault conditions. Full line-to-line rated arresters must therefore be used on the system.

Lightning Surges

In addition to the switching surges, plant equipment must also be protected from lightning surges. Lightning surges may be impressed on the plant distribution system in three ways: direct strokes, induced strokes, and by surges that are transferred through transformers. The most severe surges are caused by lightning striking the plant overhead circuits directly. The stroke current flowing in the high surge impedance of the line produces high surge voltages.
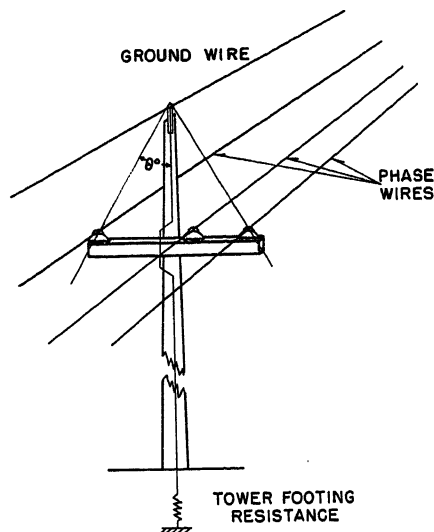
Fig. 7. Shielding of line by overhead ground wire

When lightning strikes near a line, a surge voltage is induced in the line. These surge voltages are called induced strokes. Such voltages are limited in magnitude and are usually of negligible importance on high-voltage systems. On low-voltage circuits such as the usual plant distribution circuits, however, they are important.

The third type of lightning surge that can be impressed on the plant system is that which originates on the high-voltage system and is transferred through the supply transformer to the low-voltage plant circuits. This transfer is accomplished by electrostatic and electromagnetic coupling in the transformer. This coupling can be illustrated by the simplified circuit of Fig. 5. A transformer can be represented as having two windings with a certain turns-ratio and also having distributed capacitance both to ground and between windings. For this illustration assume that the distributed capacitances are lumped into three sections; one section from the high voltage to ground C_H; one from the low voltage to ground C_L; and a mutual capacitance between windings C_M. With this circuit it can be seen that the capacitances act like an electrostatic voltage divider for a surge originating on the high-voltage side of the winding.

Fig. 5 also suggests a method of eliminating this electrostatic coupling or the voltages resulting from this coupling. If an additional capacitor or low-impedance resistor is added in parallel with C_L, the voltage ratio of the divider would be reduced to a value where it would not produce surges on the low-voltage system that would endanger the equipment. It has been determined that a 50-ohm resistance or a 0.005-microfarad capacitance

is adequate for this application.[5] It should be noted that cable circuits or other equipment connected to the transformer secondary usually provide this necessary capacitance or resistance value. For example, a total length of 50 feet of cable is usually enough to supply the necessary 0.005-microfarad capacitance. Fig. 5 also indicates an electromagnetic coupling source through the turns-ratio of the windings. This coupling source is further described in Fig. 6.[5] The actual circuit under consideration is shown in the upper right-hand corner of the figure and consists of a machine, a transformer, and a lightning arrester connected to the high-voltage side of the transformer. Directly under this circuit is shown its analogue. The machine is represented by its surge impedance R; the transformer by its leakage inductance L; the capacitance of the secondary circuit by C; and the arrester discharge voltage referred to the secondary circuit by NKE_a. The voltage impressed on the machine terminals V_m is therefore dependent on these parameters.

To determine V_m, the factors M and T must be calculated, as shown in Fig. 6, from the R, L, and C of the circuit. NKE_a can be determined from N, the ratio of the machine to the line voltages; from K, the factor given in the figure dependent on the transformer bank connection; and from E_a, the discharge voltage of the arrester. Knowing NKE_a, M, and T, the magnitude and wave shape of the voltage V_m applied to the machine can then be determined by the curves of Fig. 6. A more complete description of this method of calculation is given in reference 5.

In many cases the transferred surges determined by the foregoing method are found to be low enough so that additional protection on the secondary circuit is not needed. However, if large or important equipment is involved, the risk of failure from these surges usually justifies the small expense of the additional protective equipment.

In applying the prevention-plus-control program to lightning surges, it is found that in this case the prevention portion is the most difficult. By adequate shielding of the overhead lines on the system, direct strokes can be eliminated to a great extent; but no means have been devised to prevent effectively induced strokes or surges that are transferred through transformers.

Strictly speaking, it is impossible to eliminate all direct strokes unless the line is completely surrounded with a grounded metal enclosure. However, by proper shielding with rods, masts, or overhead ground wires, it is possible to prevent all but 0.1 per cent of the strokes to strike the shielded conductor. Fig. 7 shows an example of the overhead ground wire. For 0.1-per-cent shielding the line conductor should be within a 30-degree shielding angle from the shield wire. In addition, to prevent flashovers from ground wire to conductor, the tower ground resistance must be low and the spacing between the grounded conductor or structure and the line conductor must be sufficient, usually on the order of several feet.

Even if the overhead ground wire prevents practically all of the direct strokes, it will not prevent the appearance of lightning surges on the line conductors. When the ground wire is struck, voltages appear on the line wires by induction. These voltages plus the induced strokes and the surges that can appear through the supply transformers all can be of sufficient magnitude to damage apparatus on the plant system. Some means of control therefore is necessary to protect this apparatus.

INSULATION LEVELS

Before the control portion of the protection program is attempted, the insulation levels of the equipment to be protected must be known. Normally the equipment to be protected on an industrial system consists of cable, oil-insu-

Table I. Impulse Insulation Levels of Low-Voltage Equipment, 1½ × 40-Wave Kilovolt Crest Values

Insulation Class, Kv	Cable*	Oil-Insulated Transformer†	Dry-Type Transformer	Switchgear	Rotating Machine, Kv‡
2.5	94	45	20	60	10.25 (2.4)
5.0	113	60	25	60	16.5 (4.16)
8.66	169	75	35	75– 95§	26.1 (6.9)
15	240	95	50	95–110	50.8 (13.8)

* Recommended standards by Halperin and Shanklin.[5]

† Distribution class equipment 500 kva or less.

‡ Machine impulse insulation level is 25 per-cent higher than the 60-cycle test value. Values in the table refer to the machine ratings given in the parenthesis.

§ Switchgear value is for 7.2-kv class equipment.
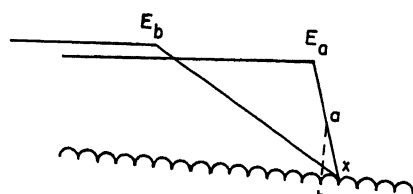
lated transformers, dry-type transformers, switchgear, and rotating machines. The impulse levels of this equipment in the 2.5- to 15-kv range are given in Table I.

The impulse levels of low-voltage overhead lines vary over a wide range but, in general, the values are higher than for cable circuits. Therefore, it is recommended that lightning arresters should be provided at the potheads when cable circuits are fed from overhead lines. In some cases arresters may be needed at the far end of these cable circuits if the cable length is excessive. This should be done to prevent the build-up of voltage at the far end resulting from reflected travelling waves in the cable circuit.

The impulse levels of oil-insulated transformers given in Table I show that protection can be provided by installing standard lightning arresters of the proper rating at the terminals of the transformer. As a general rule, these arresters should be mounted as close as possible to the terminals of the unit although, in certain cases, it is permissible to allow a specified separation distance.[6]

Dry-type power transformers and rotating machines have impulse levels lower than the other equipment. These units require special protection methods.[7] Arresters with a low protective ratio should be installed on the terminals of the transformer and, if it is connected to an exposed overhead line, then additional arresters, preferably of the expulsion type, should be installed on the line approximately 500 feet from the transformer.

The impulse strength of switchgear equipment is comparable with the strength of oil-insulated transformers. The protection of this equipment is therefore provided by standard arresters mounted in the equipment.

Table I shows the impulse strength of rotating machines to be a single value and of the lowest magnitude of all the equipment considered. In a way, this figure is misleading. While the insulation level is low compared to the other equipment, it is not a single valued quantity. The value given in the table is the conductor-to-ground insulation level but, in addition, there is a turn-to-turn insulation level that must not be exceeded. It is this

Fig. 9 (right). Means of
obtaining series inductance in
machine protection scheme
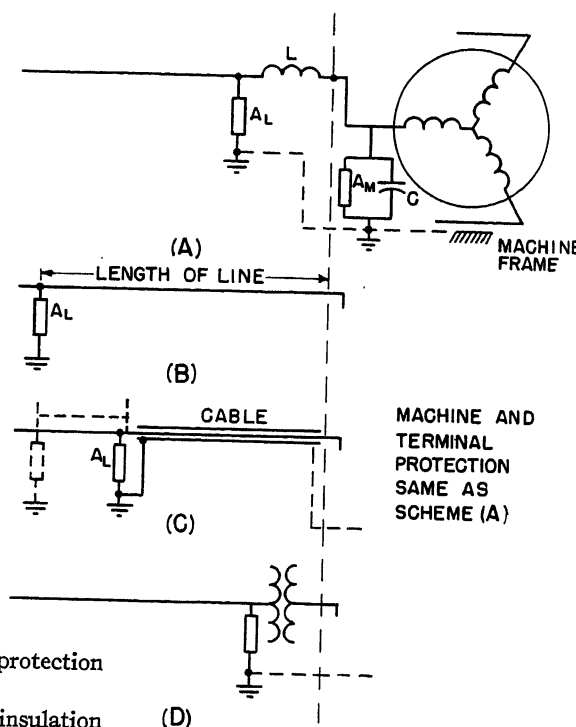
value that complicates the protection problem for rotating machines.[5]

If only the line-to-ground insulation level were to be considered, protection of the machine would be accomplished by simply applying an arrester at its terminals. Fig. 8 shows what would happen in this case to the voltage stresses between turns of the machine winding. If the machine winding could be thought of as a transmission line with distributed constants, then the voltage surge transmitted past the arrester would appear as Eₐ travelling through the winding. Fig. 8 shows that for this surge the voltage between turns would be equal to ab, which is approximately one-half the total value of the applied surge. This value of ab would in most cases be entirely too high for the turn-to-turn insulation of the machine.

Since from Fig. 8 it can be seen that the voltage between turns depends on the rate of rise of the applied surge, a control scheme becomes immediately apparent. If the surge front was sloped off, as shown by Eᵦ in Fig. 8, for the same magnitude of surge and therefore stress on the insulation to ground, the turn-to-turn stress can be reduced. To accomplish this sloping effect, the basic circuit used consists of inductance and capacitance in series. This, in effect, means applying an inductance in series with the machine and a shunt capacitor across the machine terminals. This controls the slope and, by

applying arresters across the capacitor and across the line at the junction of the line and inductance, the magnitude of the surge voltage will be controlled.

The values and form of the inductance and capacitance depend on the origin of the surge, characteristics of the supply circuit, the machine constants, and the method of grounding the machine neutral. Fig. 9 shows the most common means of providing these elements. The inductance may be a lumped choke coil or reactor as in Fig. 9(A); a section of line or cable as in Figs. 9(B) and (C); or it may be a transformer as in Fig. 9(D). In any case, the value of inductance should be at least 175 microhenrys if the line arrester is of the valve type, or 90 microhenrys if the line arrester is of the expulsion type.[5] The value of capacitance depends on the voltage rating of the machine and the method of neutral grounding. In all cases, the arrester and capacitor ground connections should be common and connected to the machine frame.

EFFECT OF NEUTRAL GROUNDING

Since a surge entering a machine winding eventually reaches the neutral of the machine, the method of neutral grounding must be considered in the machine-protection problem. If the neutral is solidly
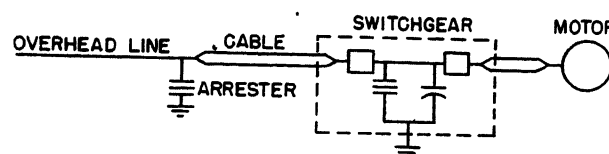
Fig. 10. Example of protection scheme

grounded, it is obvious that no surge voltage can ever exist at this point. If the neutral point is open however, under certain conditions, voltages in excess of the applied surge value are possible caused by surge reflections at the open neutral point. For this condition, some means must be provided to limit the voltages at the neutral. The most obvious method would be to provide additional arresters and capacitors at the neutral point. The neutral point is not always available, however, so another means of limiting this voltage is the further sloping of the applied wave. This is done by the addition of another set of surge capacitors at the machine terminals.

Since ungrounded machines need additional surge protection and solidly grounded machines do not, there must be a dividing line somewhere in between above which additional protective equipment is needed. It has been determined that this dividing line occurs for resistance grounding when the resistor ohms exceed one-third of the surge impedance of the machine winding. For reactance grounding, this line occurs when the 60-cycle reactance exceeds 0.2 ohm per 100 ohms of machine surge impedance. These values then determine whether a machine is grounded or ungrounded from a surge standpoint.

As mentioned before, the value of capacitance to use depends on whether a machine is grounded or ungrounded. Now that the definition of grounded operation is determined, the recommended capacitance values can be given. For all machines rated 650 volts or less, the capacitance per phase should be 1.0 microfarad. For all machines above 650 volts and less than 11.5 kv, the capacitance per phase should be 0.5 microfarad. For ungrounded machines over 11.5 kv without neutral protection, the capacitance per phase should be 0.5 microfarad. For grounded machines over 11.5 kv, the capacitance per phase should be 0.25 microfarad.

SPECIAL CONSIDERATIONS

The preceding discussion covered protection from a more or less individual piece of equipment point of view. It was demonstrated how one piece of equipment should be protected if it alone was being considered. Of course, on an actual system, different types of equipment or several pieces of the same type of equip-

ment may be located near enough to each other so that all of the units could be protected by the same protective device. An example of such a condition would be a cable feeder terminating in a switchgear cubicle from which another short cable feeder supplies a motor. Such a circuit is shown in Fig. 10. When the individual pieces of the circuit are considered, lightning arresters are needed at both ends of the cable as well as one for the switchgear, and an arrester and capacitor are needed at the motor. As shown in the figure, only two arresters and a capacitor are needed to provide this protection. One arrester at the junction of the overhead line and the cable protects the one end of the cable and the other arrester and the capacitor protect the motor, the switchgear, and the other end of the cable.

A similar situation arises when there are several motors connected to the same bus. In many cases adequate protection is obtained by mounting one set of lightning arresters and capacitors at the common bus. A conservative figure for the maximum separation distance between the motor and the protective equipment is 200 feet of line or 100 feet of cable. In most cases, greater distances would probably be satisfactory but, for important motors, these figures should probably be adhered to.

The foregoing recommendations applied primarily to equipment subjected to direct or induced strokes or surges directly attributed to such strokes. Thus, any plant system that has overhead lines exposed to direct or induced strokes would be governed by these recommendations. For a plant system that has no such exposure and contact, either directly or through a transformer, with such an exposed circuit, surge-protective equipment is not needed; e.g., a system having its own generation and no tie with a utility and the circuits of which are all cable. For this system, all that is needed is the proper grounding to prevent switching surges.

The third type of system is a combination of the preceding two. This system consists of all unexposed cable circuits but a utility tie is maintained through a step-down transformer bank. Thus surges can be transmitted through the transformer to the plant system. As mentioned previously, if the high-voltage winding of the transformer is protected by lightning arresters, the magnitude of the transferred surges are quite nominal

and, in many cases, especially when the system consists of a considerable amount of cable, additional protective equipment on the low-voltage plant system is not needed. A computer study of just this situation was made recently for a large industrial plant. The results of this study indicated that, while arresters might not be needed at all, complete protection would be assured by installing one set of arresters on the low-voltage terminals of the transformer.

Conclusions

Any program of surge protection must provide for switching surges as well as lightning surges. Switching surge protection consists of proper system grounding to prevent the occurrence of switching surges or transient overvoltages caused by arcing grounds. Lightning surge protection consists of adequate shielding to prevent direct strokes to the system and the use of lightning arresters and surge capacitors to control the surges that occur because of induced strokes and surges transmitted through transformers. By taking proper measures to provide this protection, equipment failures and plant outages resulting from these failures will be kept to a minimum.

References

1. ELECTRICAL TRANSMISSION AND DISTRIBUTION REFERENCE BOOK. Westinghouse Electric Corporation, East Pittsburgh, Pa., 4th ed., 1950, pp. 511–21.

2. APPLICATION GUIDE FOR THE GROUNDING OF SYNCHRONOUS GENERATOR SYSTEMS, AIEE Committee Report. AIEE Transactions, vol. 72, pt. III, June 1953, pp. 517–30.

3. KLYDONOGRAPH SURGE INVESTIGATIONS, J. H. Cox, P. H. McAuley, L. G. Huggins. AIEE Transactions, vol. 46, Feb. 1927, pp. 315–29.

4. LIGHTNING INVESTIGATION ON TRANSMISSION LINES—II, W. W. Lewis, C. M. Foust. AIEE Transactions, vol. 50, Sept. 1931, pp. 1139–46.

5. LIGHTNING PROTECTION FOR ROTATING MACHINES, G. D. McCann, E. Beck, L. A. Finzi. AIEE Transactions (Electrical Engineering), vol. 63, 1944, pp. 319–33.

6. SURGE PROTECTION OF TRANSFORMERS BASED ON NEW LIGHTNING-ARRESTER CHARACTERISTICS, J. K. Dillard, T. J. Bliss. AIEE Transactions, vol. 73, pt. III-B, Oct., 1954, pp. 1305–11.

7. DIELECTRIC STRENGTH AND PROTECTION OF MODERN DRY-TYPE AIR-COOLED TRANSFORMERS, P. L. Bellaschi, E. Beck. AIEE Transactions (Electrical Engineering), vol. 64, Nov. 1945, pp. 759–62.

8. IMPULSE STRENGTH OF INSULATED-POWER-CABLE CIRCUITS, Herman Halperin, G. B. Shanklin. AIEE Transactions, vol. 63, 1944, p. 1190–97.

9. APPLICATION GUIDE ON METHODS OF NEUTRAL GROUNDING OF TRANSMISSION SYSTEMS, AIEE Committee Report. Ibid., Aug. 1953, pp. 663–68.

Discussion follows

Discussion

Newell A. Williams (The Goodyear Tire and Rubber Company, Akron, Ohio): Mr. Wagner's paper was of great interest to me because of some experiences that we had with motor failures due to surges several years ago. During a storm, lightning struck a ventilation stack at one of our large tire plants. Coming down the stack, the bolt jumped to the frame of a 400-horsepower 4,600-volt mill motor. Evidently the frame of the motor was not properly grounded at the time, for the stroke penetrated the winding, causing damage to several coils and simultaneous failure of another large motor on the 4,600-volt bus section.

The burned coils were replaced on the 400-horsepower mill motor and it was put back in service. Six weeks later the same motor again failed and at the same time an 800-horsepower Banbury motor burned out. Again coils were replaced and it was put back into service. Two weeks after, it again went to ground, this time causing the simultaneous failure of another 600-horsepower Banbury motor and another 600-horsepower mill motor.

This time the motor was completely rewound. When the old winding was removed, it was found that the cell insulation was "peppered" with small holes that had been burned through it, presumably by the original lightning stroke. The insulation had broken down on two subsequent occasions and caused surges that had ruined three other large motors.

With this evidence we decided that we should ground the 4,600-volt system. Because of the expense of providing ground protection on all of our present feeders, it was decided to use the high-resistance grounding scheme illustrated by Mr. Wagner in his Fig. 4(B). This system was installed and the results have been very gratifying. To my knowledge, no multiple motor failures have occurred since.

We have installed similar systems in all of our large plants since that time and feel that this is very cheap insurance against trouble from surges.

R. H. Kaufmann (General Electric Company, Schenectady, N. Y.): Mr. Wagner's paper serves to bring into sharp focus the evils of transitory overvoltages in industrial systems. The paper refocuses attention on preventable overvoltages in industrial power systems. It points out tried and tested methods of curbing such overvoltages and emphasizes that overvoltage reduction means longer electric equipment life and fewer insulation breakdowns. He cites cases in which the application of system grounding was observed to result in a reduced rate of circuit tripouts. The discussion of the benefits of system grounding should aid in speeding universal adoption of this desirable practice.

In the treatment of resistance grounding, the reader might easily acquire an over-optimistic impression of the quality of high-resistance grounding.

While it is true that the severe overvoltages resulting from intermittent ground faults and high-inductive reactance line-to-ground circuits can be curbed by high-resist-ance grounding, it is incorrect to imply that high-resistance grounding eliminates overvoltages. There are a number of overvoltage sources, for example, a cross-connection with a higher voltage grounded system, which will not be eliminated by high-resistance grounding.

Furthermore, a high-resistance grounding system allows any one ground fault fully to displace the system neutral, thus resulting in the application of 73-per-cent overvoltage on two phases throughout the system. The absence of automatic trip on a ground fault will allow this overvoltage to be sustained. It is important to realize that adoption of high-resistance grounding should be accompanied by installation of ground-fault searching equipment to enable the ground fault to be quickly located and quite promptly cleared, lest the continued overvoltage throughout the rest of the system nullify the advantage of avoiding an immediate tripout on the fault member.

The presence or absence of rotating machines on the system in question is an important factor in deciding whether high-resistance grounding is likely to be a good choice. Rotating-machine winding faults incur high short-circuit currents around the short-circuited turns, in spite of a negligible low ground current. The high turn-short-circuit current liberates much heat and melts copper and core laminations. The absence of substantial ground current makes machine differential relaying inoperative until the fault burns into another phase and also makes ground overcurrent relaying ineffective. The high-resistance system grounding sacrifices the possible more sensitive protective relaying which would be obtained with low-resistance reactance, or solid grounding. This sacrifice may mean longer, more costly, motor-winding repair on those motors which develop internal winding failure.

Under "Resistance Grounding," Mr. Wagner states "R_0/X_0 should always be equal to or greater than 2, and X_0/X_1 must be less than 20." It is believed that only the first limit was intended; namely, R_0/X_0 should be equal to or greater than 2. The use of distribution transformers for the grounding circuit, as suggested in Fig. 4, is in itself, evidence that X_0/X_1 ratios higher than 20 are accepted.

Under "Reactance Grounding," Fig. 1 plots system overvoltage as a function of X_0/X_1 ratio. System overvoltages are not the direct result of a given X_0/X_1 ratio, but rather the result of some particular disturbing influence in the presence of the stated X_0/X_1 value. Furthermore, the voltage magnitudes will be affected by other system parameters, such as X_{c0}/X_{c1} and R_0/X_0. To allow proper appraisal of Fig. 1, Mr. Wagner should define the exact character of the disturbing influence, as well as the values of other pertinent circuit parameters. Perhaps, this curve was extracted from published literature in which case a reference thereto would suffice.

Under "Lightning Surges" is expressed the thought that by proper shielding of open-wire circuits, freedom from breakdown from direct lightning strokes can be attained except for perhaps one case in a thousand. To achieve such levels of performance requires high insulation levels and wide conductor spacings, such as might be found on lines operating at 230 kv,

or higher. For industrial system circuits operating generally at voltages below 15 kv, the afore-mentioned performance levels rarely can be even approached. This comment is not to be construed to mean that shielding should be abandoned, but rather that we should know and understand the facts as they are.

It is my earnest intent that these comments serve to enrich and enlarge the interpretations of this informative paper.

C. L. Wagner: Mr. Williams' and Mr. Kaufmann's comments on this paper demonstrate very clearly the true worth of discussions on technical papers. In the first place, the logical question to ask after reading or hearing any paper giving general application recommendations on a given subject is "Where have those recommended practices been used and what success has been obtained?" Mr. Williams' discussion supplies the answer to this question, at least for the section of the paper pertaining to high-resistance grounding.

Mr. Kaufmann's comments illustrate the complexities involved in preparing a general application paper on as broad a subject as surge protection on industrial systems. It is extremely difficult to cover all the factors in the problem in sufficient detail to insure comprehension by the reader and at the same time limit the paper to a reasonable length. Further discussion on points not sufficiently clear in the formal presentation is certainly worth-while.

In this regard the author would first like to make clear that no particular type of system grounding was recommended in this paper. All that was stated or intended in this paper was that by proper system grounding, the occurrence of transient overvoltage due to switching surges or arcing ground faults could effectively be prevented. I have attempted to define the rules that have been established for all types of system grounding to limit the magnitude of these switching surge voltages.

Mr. Kaufmann evidently feels that too much emphasis was placed on high-resistance grounding. No such emphasis was intended. There are too many factors to be considered in selecting a particular type of grounding scheme to include a discussion of same in this paper. References 1, 2, and 3 of the paper are but several of the many works written on the subject, which discuss the relative merits of the various schemes.

Mr. Kaufmann states that the paper implies that high-resistance grounding eliminates overvoltages. I never stated or certainly meant to imply that it does so. It eliminates transient overvoltages as stated but not overvoltages. In fact, no type of system grounding, even solid grounding, will eliminate the overvoltage occurring in the example given by Mr. Kaufmann of a high-voltage line falling across a lower voltage line. The ground relays might trip the circuit if the system were solidly grounded but the overvoltage would still be impressed on the system for a short time.

The need for or advantages of ground-detecting relays or alarms on high-resistance grounded systems is further emphasized by Mr. Kaufmann's comments. The presence of line-to-line voltage impressed on the line-to-ground insulation of plant equipment will reduce the life of the insulation to some

degree so that it is advantageous to clear the fault as soon as possible.

I do not understand fully Mr. Kaufmann's comments on a fault in an ungrounded machine. If a single ground fault occurs in the machine, there will be no turns short circuited and thus no high circulating currents. The only current causing damage in the machine will be that limited by the high grounding resistance. This current will normally cause a negligible amount of damage. If there is a turn-to-turn fault in the machine not involving ground, then there will be high circulating currents in the faulted turn but these currents will not be detected by the differential relays whether the system is solidly grounded or ungrounded. The only condition the author can visualize where Mr. Kaufmann's comments would apply, is when the turn-to-turn fault finally burns to ground. For this condition, most certainly the machine on the solidly grounded system would be relayed off whereas the machine on the high-resistance grounded system would not be tripped until the fault involved another phase. It is my opinion, however, that in either case the damage in the machine would be extensive and require major repairs.

Without going into further details on the relative merits of various grounding schemes which is actually beyond the scope of the paper, I believe Mr. Kaufmann would agree simply to state that if plant operating procedures definitely prohibit circuit tripping on ground faults, then high-resistance grounding is the best scheme to use. If circuit tripping is permissable, then some form of low-impedance grounding is preferable.

Mr. Kaufmann raised the question as to the circuit conditions involved when the data of Fig. 1 were obtained. This is a very good point and definite references for it should have been stated in the paper. Rather than give all the pertinent data at this time, which is actually not necessary for this discussion, the author would like to refer those interested to page 520 of reference 1 of the paper. As Mr. Kaufmann states, this curve will not necessarily be exactly correct for all values of circuit parameters and switching conditions. It was included in the paper to show first of all that the magnitudes of switching surges can be quite high (above 500 per cent of normal) if proper precautions are not taken. It was also used to show that without forcing current zero in the arc path (one of the conditions under which the data were taken) there is a sharp break in the curve around the value of X_0/X_1 equals 10. This demonstrates where the rule of X_0/X_1 equal to or less than 10 came from.

The question of X_{c0}/X_{c1} and R_0/X_0 ratios used in these studies leads directly into the answer to the question of where did the statement "R_0/X_0 should always be equal to or greater than 2, and X_0/X_1 must be less than 20" come from. Without going into any detail, it might be said that recent analogue computer studies have shown that for certain X_{c0}/X_{c1} and X_{c1}/X_1 ratios it is not sufficient to merely keep R_0/X_0 greater than 2. It has been found that the added restriction of X_0/X_1 less than 20 must be used to make sure that for any system capacitance values the transient overvoltages will not be excessive.

The last comment of Mr. Kaufmann's is a misinterpretation of the author's statement. The paper states "by proper shielding . . ., it is possible to prevent all but 0.1 per cent of the strokes to strike the shielded conductor." This refers only to shielding of the conductors from direct strokes and not to back flashes or flashovers from shield wire to conductor as Mr. Kaufmann interpreted it. Mr. Kaufmann is correct in that it is impossible to limit back flashes to such a low level on low-voltage lines. In fact even on 230 kv and above, it is impossible to obtain this level.

I would like to thank the discussers for their comments which no doubt added immeasurably to the worth of this paper.

Simple Analytic Method for Linear Feedback System Dynamics

M. E. CLYNES
NONMEMBER AIEE

THE Laplace transform method for analyzing the transient response of feedback systems is laborious.[1] Synthesis by this method is almost impracticable. The graphical methods which were developed by Nyquist, Bode, Evans, and others allowed the analysis and synthesis of linear feedback systems to be made with considerably less effort.[2,3,4] Methods using infinite series to approximate the transient response from the open-loop transfer function were developed by Tustin, and earlier by Bedford and Fredendall.[5,6]

The analytic method presented here takes as its starting point the open-loop transfer function. It eliminates the calculation, in each individual case, of the

closed loop transfer function and the solution of the characteristic equation. General solutions of the characteristic equations are given in terms of the open-loop parameters, for the systems considered in this paper, viz., single-loop systems and simple multiple-loop systems. These solutions give the transient behavior.

The method is well suited for synthesis, as the characteristics of the system components keep their identity in the analysis. The results are accurate for certain systems, while for the others approximate solutions are given well within most practical requirements.

Nomenclature

K, K_v, K_a = steady-state open-loop gain of 0-type, 1-type, and 2-type systems respectively

$\omega_1, \omega_2, \omega_3$ = reciprocals of τ_1, τ_2, τ_3, the open-loop lag time constants respectively

ω = the undamped natural frequency of the closed-loop system

ζ = the damping factor of the closed-loop response, i.e., of the governing complex roots

ζ_{III} = the damping factor of our imaginary system having the three slowest lags of the system considered with the same system gain

ζ_{II} = the damping factor of an imaginary system having the two slowest lags of the system considered, with the same system gain

τ = the pure exponential decay time constant in 3-lag system transients

n = the gain margin for uncompensated 3-lag systems

ρ = the ratio of the multipliers of the two transient components A/B

ψ = angle

2-Time Lag System

As an introduction to the method, consider a system of an open-loop transfer function $G(s)$. The closed-loop transfer function is then

$$M(s) = \frac{G(s)}{1+G(s)} \tag{1}$$

Consider first

$$G = \frac{K}{(1+\tau_1 s)(1+\tau_2 s)}$$

Then

$$M(s) = \frac{K}{K+(1+\tau_1 s)(1+\tau_2 s)}$$

$$= \frac{\dfrac{K}{K+1}}{1+\dfrac{\tau_1+\tau_2}{K+1}s+\dfrac{\tau_1\tau_2}{K+1}s^2}$$

The denominator, a quadratic in s, is of the form

$$1+\frac{2\zeta}{\omega}s+\frac{s^2}{\omega^2}$$

which represents a pair of complex roots of ζ and ω corresponding to the transient

$$\frac{1}{\sqrt{(1-\zeta^2)}}\,\epsilon^{-\zeta\omega t}\times$$

$$\sin\left(\omega\sqrt{(1-\zeta^2)}t-\tan^{-1}\frac{\sqrt{(1-\zeta^2)}}{\zeta}\right)$$

for a step displacement input.

Comparing coefficients of like powers of s

$$\frac{2\zeta}{\omega}=\frac{\tau_1+\tau_2}{K+1}\qquad(2)$$

and

$$\frac{1}{\omega^2}=\frac{\tau_1\tau_2}{K+1}\qquad(3)$$

giving

$$\omega^2=\omega_1\omega_2(K+1)\qquad(4A)$$

$$\zeta=\frac{\omega_1+\omega_2}{2\omega}\qquad(4B)$$

The solution for ω and ζ is simple in this case as the equations 2, and 3 are quadratic in ω and ζ.

3-Time Lag System

The difficulties in applying the same method to a 3-lag system are overcome by the convenient introduction of a new parameter as shown in the following:

Consider a system of open-loop transfer function

$$\frac{K}{(1+\tau_1s)(1+\tau_2s)(1+\tau_3s)}$$

The closed-loop transfer function is

$$M(s)=$$

$$\frac{\dfrac{K}{K+1}}{1+\dfrac{\tau_1+\tau_2+\tau_3}{K+1}s+\dfrac{\tau_1\tau_2+\tau_2\tau_3+\tau_3\tau_1}{K+1}s^2+\dfrac{\tau_1\tau_2\tau_3}{K+1}s^3}$$

The denominator is of the form

$$(1+\tau s)\left(1+\frac{2\zeta}{\omega}s+\frac{s^2}{\omega^2}\right)$$

Here the two factors correspond to the two components of the transient response $A\epsilon^{-t/\tau}$ and $B\epsilon^{-\zeta\omega t}\sin[\omega\sqrt{(1-\zeta^2)}t+\psi]$. A, B, and the angle ψ depend on the type of disturbance.

Equating coefficients of the like powers of s

$$\tau+\frac{2\zeta}{\omega}=\frac{\tau_1+\tau_2+\tau_3}{K+1}\qquad(5)$$

$$\frac{2\tau\zeta}{\omega}+\frac{1}{\omega^2}=\frac{\tau_1\tau_2+\tau_2\tau_3+\tau_3\tau_1}{K+1}\qquad(6)$$

and

$$\frac{\tau}{\omega^2}=\frac{\tau_1\tau_2\tau_3}{K+1}\qquad(7)$$

A solution of these equations for any of the unknowns ω, ζ, τ would involve solving a cubic equation. However, let

$$\tau=\frac{1}{n}\frac{\tau_1+\tau_2+\tau_3}{K+1}\qquad(8)$$

where n is a function of τ_1, τ_2, τ_3, and K. Then

$$\frac{2\zeta}{\omega}=\frac{\tau_1+\tau_2+\tau_3}{K+1}\frac{n-1}{n}\qquad(9)$$

Substituting in equation 6 and using equation 7 to eliminate $1/\omega^2$, the following equation containing only n as unknown is obtained

$$\frac{n-1}{n^2}\frac{(\tau_1+\tau_2+\tau_3)^2}{(1+K)^2}+\frac{n}{n}\frac{\tau_1\tau_2\tau_3}{\tau_1+\tau_2+\tau_3}$$
$$=\frac{\tau_1\tau_2+\tau_2\tau_3+\tau_3\tau_1}{K+1}\qquad(10)$$

The first term on the left-hand side may be neglected with a proportional error of

$$\epsilon=\frac{n-1}{n^3}\frac{(\tau_1+\tau_2+\tau_3)}{\tau_1\tau_2\tau_3(K+1)^2}$$

The magnitude of this error will be evaluated in the following. Then

$$n=\frac{(\tau_1+\tau_2+\tau_3)(\omega_1+\omega_2+\omega_3)}{K+1}\qquad(11)$$

and, since the term $2\tau\zeta/\omega$ in equation 6 is neglected

$$\omega^2=(K+1)\frac{\omega_1\omega_2\omega_3}{\omega_1+\omega_2+\omega_3}\qquad(12)$$

From equation 7

$$\tau=\frac{1}{\omega_1+\omega_2+\omega_3}\qquad(13)$$

and from equation 5

$$\zeta=\frac{n-1}{2}\frac{\omega}{\omega_1+\omega_2+\omega_3}\qquad(14)$$

The proportional error ϵ resulting by omitting term $2\tau\zeta/\omega$ compared to $1/\omega^2$ in equation 6 is

$$2\tau\zeta\omega=\frac{(n-1)\omega^2}{(\omega_1+\omega_2+\omega_3)^2}=\frac{4\zeta^2}{n-1}\qquad(15)$$

making use of the approximate relations of equations 11, 12, 13, and 14.

The corrected value of ω^2 is therefore

$$\omega'^2=\frac{\omega^2}{1-\epsilon}\qquad(16)$$

giving the following corrected values

$$\tau'=\frac{\tau}{1-\epsilon}\qquad(17)$$

and

$$\zeta'=\zeta\frac{\left(1-\dfrac{\epsilon}{(n-1)(1-\epsilon)}\right)}{\sqrt{(1-\epsilon)}}=\zeta\left(1+\frac{\epsilon}{2}-\frac{\epsilon}{n-1}\right)\qquad(18)$$

if ϵ is small and n stands for

$$\frac{(\omega_1+\omega_2+\omega_3)(\tau_1+\tau_2+\tau_3)}{K+1}$$

The accuracy of equations 12, 13, and 14 increases with spread of the time lags. In many cases ϵ is sufficiently small so that correction is not necessary. The uncorrected values error on the side of safety for all values of $n>3$, while for $n<3$ as $n\to1$ the error approaches zero.

From equation 14 it is seen that ζ becomes negative for $n<1$. Thus $n<1$ indicates instability. From equation 11 it follows that n is actually the gain margin, i.e., the factor by which $K+1$ must be increased for instability to occur. Thus, the value of n is immediately indicative of the stability of the system.

To obtain the transient response, ω and ζ are evaluated. For an uncompensated 3-time lag system, τ is short compared to $1/\omega$ and the coefficient of $\epsilon^{-t/\tau}$ in the transient is small; therefore, this term of the transient often is negligible. The transient is governed by the complex roots of the characteristic equation.

Example

$\tau_1=1$ second, $_2\tau=0.1$ second, $\tau_3=0.02$ second, $K=15$

then

$$n=\frac{61\times1.12}{61}=4.27 \text{ by equation 11}$$

$$\omega^2=\frac{16\times500}{61}=131.5 \text{ by equation 12}$$

$$\tau=\frac{1}{16}=0.0164 \text{ by equation 13}$$

$$\zeta=\frac{3.27}{2}\times\frac{11.46}{61}=0.307 \text{ by equation 14}$$

Error

$$\epsilon=\frac{(0.307)^2\times4}{3.27}=0.115 \text{ by equation 15}$$

$$\omega'^2=\frac{131.5}{1-0.115}=148.5 \text{ by equation 16}$$

$\omega'=12.2 \text{ second}^{-1}$ by equation 17

$\tau'=0.0186$ second by equation 17

$\zeta'=0.307(1+0.058-0.035)=0.314$ by equation 18

These are the same values as are obtained by the classic method of calculating the closed-loop transfer function and solving the cubic in the denominator.

Open Loop Transfer Function	Closed Loop Response
$\dfrac{K}{(1+\tau_1 s)(1+\tau_2 s)}$	$\omega^2 = \omega_1\omega_2(K+1), \quad \zeta = \dfrac{\omega_1+\omega_2}{2\omega}$
$\dfrac{K}{(1+\tau_1 s)(1+\tau_2 s)(1+\tau_3 s)}$	$\begin{cases} \omega^2 = \dfrac{\omega_1\omega_2\omega_3(K+1)}{\omega_1+\omega_2+\omega_3}* \\[4pt] n = \dfrac{(\tau_1+\tau_2+\tau_3)(\omega_1+\omega_2+\omega_3)}{K+1} \\[4pt] \zeta = \dfrac{n-1}{2}\dfrac{\omega}{\omega_1+\omega_2+\omega_3},* \quad \tau = \dfrac{1}{\omega_1+\omega_2+\omega_3}* \\[6pt] \text{Error } \epsilon = \dfrac{4\zeta^2}{n-1} \\[4pt] \text{Corrected values} \\[4pt] \omega'^2 = \dfrac{\omega^2}{1-\epsilon}, \quad \tau' = \dfrac{\tau}{1-\epsilon}, \quad \zeta' = \zeta\left(1+\dfrac{\epsilon}{2}-\dfrac{\epsilon}{n-1}\right) \end{cases}$
$\dfrac{K}{(1+\tau_1 s)(1+\tau_2 s)(1+\tau_3 s)\dots(1+\tau_n s)}$ $\tau_1 > \tau_2 > \dots > \tau_n$	$\begin{cases} \omega^2 = \dfrac{K+1}{\dfrac{1}{2}\sum\limits_{\substack{1 \\ i\neq k}}^{n}\tau_i\tau_k}* \\[10pt] \zeta_N = \zeta_{III}\left(1-\dfrac{K(\tau_4+\tau_5+\dots+\tau_n)}{\tau_1+\tau_2+\tau_3\dots+\tau_n}\right)* \end{cases}$
$\dfrac{K(1+\tau_3 s)}{(1+\tau_1 s)(1+\tau_2 s)}$	$\omega^2 = \omega_1\omega_2(K+1), \quad \zeta = \dfrac{1}{2}\left(\dfrac{K\tau_3}{\tau_1\tau_2\omega}+\dfrac{\omega_1+\omega_2}{\omega}\right)$
$\dfrac{K(1+\tau_4 s)}{(1+\tau_1 s)(1+\tau_2 s)(1+\tau_3 s)}$	$\begin{cases} \omega^2|\zeta_{max} = \dfrac{3(K+1)\omega_1\omega_2\omega_3}{\omega_1+\omega_2+\omega_3} \\[6pt] \zeta_{max} = \dfrac{\omega_1+\omega_2+\omega_3}{3\omega} \\[6pt] \tau_4|\zeta_{max} = \dfrac{1}{K}\left(\dfrac{3(K+1)}{\omega_1+\omega_2+\omega_3}+\dfrac{2}{3}\dfrac{(K+1)(\omega_1+\omega_2+\omega_3)}{\omega^2}-(\tau_1+\tau_2+\tau_3)\right) \end{cases}$
$\dfrac{K_v}{s(1+\tau_1 s)}$	$\omega^2 = \dfrac{K_v}{\tau_1}, \quad \zeta = \dfrac{\omega}{2K_v}$
$\dfrac{K_v}{s(1+\tau_1 s)(1+\tau_2 s)}$	$\begin{cases} \omega^2 = \dfrac{K_v}{\tau_1+\tau_2},* \quad n = \dfrac{\omega_1+\omega_2}{K_v} \\[4pt] \zeta = \dfrac{n-1}{2}\dfrac{\omega}{\omega_1+\omega_2},* \quad \tau = \dfrac{1}{\omega_1+\omega_2}* \\[6pt] \text{Error } \epsilon = \dfrac{4\zeta^2}{n-1} \\[4pt] \text{Corrected values} \\[4pt] \omega'^2 = \dfrac{\omega^2}{1-\epsilon}, \quad \tau' = \dfrac{\tau}{1-\epsilon} \\[4pt] \zeta' = \zeta\left(1+\dfrac{\epsilon}{2}-\dfrac{\epsilon}{n-1}\right) \end{cases}$
$\dfrac{K_v}{s(1+\tau_1 s)(1+\tau_2 s)\dots(1+\tau_n s)}$	$\begin{cases} \omega^2 = \dfrac{K_v}{\tau_1+\tau_2+\dots+\tau_n}* \\[4pt] \zeta_n = \zeta_{II}[1-K_v(\tau_3+\tau_4+\dots+\tau_n)]* \end{cases}$
$\dfrac{K_v(1+\tau_2 s)}{s(1+\tau_1 s)}$	$\omega^2 = \dfrac{K_v}{\tau_1}, \quad \zeta = \dfrac{\omega}{2}\left(\tau_2+\dfrac{1}{K_v}\right)$
$\dfrac{K_v(1+\tau_3 s)}{s(1+\tau_1 s)(1+\tau_2 s)}$	$\begin{cases} \omega^2|\zeta_{max} = \dfrac{3K_v}{\tau_1+\tau_2}, \quad \zeta_{max} = \dfrac{\omega_1+\omega_2}{3\omega} \\[6pt] \tau_3|\zeta_{max} = \dfrac{3}{\omega_1+\omega_2}+\dfrac{2}{3}\dfrac{\omega_1+\omega_2}{\omega^2}-\dfrac{1}{K_v} \end{cases}$
$\dfrac{K_a(1+\tau_2 s)}{s^2(1+\tau_1 s)}$	$\begin{cases} \omega^2|\zeta_{max} = 3K_a, \quad \zeta_{max} = \dfrac{\omega_1}{3\omega} \\[6pt] \tau_2|\zeta_{max} = \dfrac{3}{\omega_1}+\dfrac{2}{3}\dfrac{\omega_1}{\omega^2} \end{cases}$

The transient response may then be evaluated completely in the standard way for any given type of input. For a step displacement input, the output is

$$\frac{K}{K+1}\left[1-\frac{2\tau^2\omega^2}{1-2\tau\zeta\omega+\tau^2\omega^2}e^{-t/\tau}+\right.$$
$$\left. e^{-\zeta\omega t}\frac{\sin[\omega\sqrt{(1-\zeta^2)}t-\psi]}{\sqrt{(1-\zeta^2)}\sqrt{(1-2\tau\zeta\omega+\tau^2\omega^2)}}\right]$$

where

$$\psi = \tan^{-1}\frac{\sqrt{(1-\zeta^2)}}{-\zeta} + \tan^{-1}\frac{\sqrt{(1-\zeta^2)}\tau\omega}{1-\tau\zeta\omega}$$

This is generally not necessary, however, since a knowledge of ω, τ, and ζ fix the main characteristics of the transient for a 3-lag system. The design is based on a desired value of ζ and ω.

Extension to n-Time Lag System

Comparison of ω of the 2- and 3-time lag systems for the same gain and with the same time lags τ_1 and τ_2 leads to a formulation of a more general expression. Let the time lags be arranged in order of magnitude so that τ_1 is the longest. Then, if Roman numerals indicate the number of lags in the system

$$\omega_{II}^2 = \frac{K+1}{\tau_1\tau_2} = (K+1)\omega_1\omega_2$$

$$\omega_{III}^2 = \frac{K+1}{\tau_1\tau_2+\tau_2\tau_3+\tau_3\tau_1} = \frac{(K+1)\omega_1\omega_2\omega_3}{\omega_1+\omega_2+\omega_3} \quad (19)$$

$$\omega_{IV}^2 = \frac{K+1}{\tau_1\tau_2+\tau_1\tau_3+\tau_1\tau_4+\tau_2\tau_3+\tau_2\tau_4+\tau_3\tau_4}$$

$$= \frac{(K+1)\omega_1\omega_2\omega_3\omega_4}{\omega_1\omega_2+\omega_1\omega_3+\omega_1\omega_4+\omega_2\omega_3+\omega_2\omega_4+\omega_3\omega_4}$$

$$= \omega_{III}^2 \frac{1}{1+\dfrac{\tau_4(\tau_1+\tau_2+\tau_3)}{\tau_1\tau_2+\tau_2\tau_3+\tau_3\tau_1}}$$

since τ_4 is the smallest time lag. To obtain a more correct value for ω_{IV}, the corrected value of ω_{III} can be used. In more general form

$$\omega_N^2 = \frac{1}{2}\frac{K+1}{\sum\limits_{\substack{i,j=1 \\ i\neq j}}^{n}\tau_i\tau_j} \quad (20)$$

ω_N as given by this equation can be corrected by the correction for ω_{III} for a more accurate result.

It is seen that the addition of further shorter time lags decreases ω only slightly, and ω is largely determined by the three slowest lags in the system.

The equations for the n-lag systems are approximate. They have been verified on the analogue computer and have proved to be useful. Since the analogue computer itself is approximate, a proper evaluation of the accuracy will have to form a separate study.

Although ω decreases only slightly by addition of further shorter lag terms in the open-loop transfer function, ζ is diminished considerably. A comparison, similar to the foregoing, of ζ_{II} and ζ_{III} leads to the following approximate expressions

$$\zeta_{IV} = \zeta_{III}\left(1-\frac{K\tau_4}{\tau_1+\tau_2+\tau_3}\right) \quad (21)$$

$$\zeta_N = \zeta_{III}\left(1-\frac{K\tau_4+\tau_5+\dots+\tau_n}{\tau_1+\tau_2+\tau\dots\dots\tau_n}\right) \quad (22)$$

The net effect of the changes in ζ and ω is to leave $\omega\sqrt{(1-\zeta^2)}$ effectively unchanged. Thus, the actual damped frequency of the system is practically unaffected by additional smaller lags. There

Table I (Continued)

Open Loop Transfer Function	Closed Loop Response

$\dfrac{K}{(1+\tau_1 s)(1+\tau_2 s)}$
Rate feedback
$\dfrac{HT_3 s}{1+T_3 s}$

(a) $T_3 > \dfrac{1}{\omega}$

$\omega^2 = (1+K+KH)\omega_1\omega_2$*

$\zeta = \dfrac{1}{2}\omega\left(\dfrac{\tau_1+\tau_2}{1+K+KH} + \dfrac{\tau_1\tau_2}{\tau_1}\dfrac{KH}{(1+K+KH)^2}\right)$*

$\tau = \tau_1\left(\dfrac{1+K+KH}{K+1}\right)$*

Error $\epsilon = -\dfrac{\tau_1+\tau_2}{1+K} + \dfrac{2\zeta}{\omega}$

Corrected values $\tau' = \tau + \epsilon$

For correct ω and ζ use equations (33) and (32)

(b) $T_3 < \dfrac{1}{\omega}$

$\omega^2 = \dfrac{(K+1)\omega_1\omega_2\omega_3}{\omega_1+\omega_2+\omega_3}$* $\tau = \dfrac{n'-1}{2}\dfrac{\omega}{\omega_1+\omega_2+\omega_3}$*

$n' = \dfrac{[\tau_1+\tau_2+\tau_3(1+K+KH)](\omega_1+\omega_2+\omega_3)}{K+1}$

Error $\epsilon = \dfrac{4\zeta^2}{n'-1}$

Corrected values $\omega'^2 = \dfrac{\omega^2}{1-\epsilon}$, $\zeta' = \zeta\left(1+\dfrac{\epsilon}{2}-\dfrac{\epsilon}{n'-1}\right)$

$\dfrac{K}{(1+\tau_1 s)(1+\tau_2 s)(1+\tau_3 s)}$
Rate feedback
Hs

$\omega^2|_{\zeta_{max}} = \dfrac{3(K+1)\omega_1\omega_2\omega_3}{\omega_1+\omega_2+\omega_3}$, $\tau|_{\zeta_{max}} = \dfrac{3}{\omega_1+\omega_2+\omega_3}$, $\zeta_{max} = \dfrac{\omega_1+\omega_2+\omega_3}{3\omega}$

$H|_{\zeta_{max}} = \dfrac{1}{K}\left(\dfrac{3(K+1)}{\omega_1+\omega_2+\omega_3} + \dfrac{2}{3}(K+1)\dfrac{\omega_1+\omega_2+\omega_3}{\omega^2} - (\tau_1+\tau_2+\tau_3)\right)$

$\rho = \dfrac{-\tau^2\omega^2\sqrt{1-\zeta^2}}{(1-2\tau\zeta\omega+\tau^2\omega^2)^{1/2}}$

$\dfrac{K_v}{s(1+\tau_1 s)(1+\tau_2 s)}$
Rate feedback
Hs

$\omega^2|_{\zeta_{max}} = \dfrac{3K_v}{\tau_1+\tau_2}$, $\zeta_{max} = \dfrac{\omega_1+\omega_2}{3\omega}$

$H|_{\zeta_{max}} = \dfrac{3}{\omega_1+\omega_2} + \dfrac{2}{3}\dfrac{\omega_1+\omega_2}{\omega^2} - \dfrac{1}{K_v}$

$\dfrac{K_a}{s^2(1+\tau_1 s)}$
Rate feedback
Hs

$\omega^2|_{\zeta_{max}} = 3K_a$, $\zeta_{max} = \dfrac{\omega_1}{3\omega}$, $H|_{\zeta_{max}} = \dfrac{3}{\omega_1} + \dfrac{2}{3}\dfrac{\omega_1}{\omega^2}$

* Equations are approximate.

---

is, however, an appreciable decrease in the damping and corresponding increase in overshoots.

Again, as in the 3-lag system, with higher order lag systems, the pure exponential decay terms in the transient are small and the behavior is governed by the single pair of complex roots. With single-loop multilag systems, more than one pair of complex roots do not arise regardless of how many lags there are in the system.

Derivation of Equations for Type-1 and Type-2 Systems

All the equations derived so far for type-0 systems, and those to follow, can be converted to type-1 systems, i.e., systems with an integrator in the open loop, by letting $\tau_1 \to \infty$ and substituting $K_v\tau_1$ for K where K_v is now the velocity constant for the type-1 system. The other time constants are renumbered to denote the time constants of the type-1 system. The corresponding equations are given in Table I.

For type-2 systems, i.e., systems with two integrations in the open loop, only the compensated transfer functions are of interest since such a system is always

unstable if uncompensated. Equations for such type-2 systems are derived from those of type-1 by letting $\tau_1 \to \infty$ and substituting K_a, τ_1 for K_v. K_a is the acceleration constant of the type-2 system.

Lead Compensation

One of the most common ways of increasing the stability and improving the time response simultaneously is to insert lead in the open-loop transfer function. Consider the open-loop transfer function

$$\dfrac{K(1+\tau_4 s)}{(1+\tau_1 s)(1+\tau_2 s)(1+\tau_3 s)}$$

It is desired to find an optimum value of τ_4. The corresponding closed-loop transfer function is

$$\dfrac{K}{K+1}\dfrac{1+\tau_4 s}{1+\dfrac{\tau_1+\tau_2+\tau_3+K\tau_4}{K+1}s + \dfrac{\tau_1\tau_2+\tau_2\tau_3+\tau_3\tau_1}{K+1}s^2 + \dfrac{\tau_1\tau_2\tau_3}{K+1}s^3}$$

Now, as for the 3-lag system

$$\tau + \dfrac{2\zeta}{\omega} = \dfrac{\tau_1+\tau_2+\tau_3+K\tau_4}{K+1} \qquad (23)$$

$$\dfrac{2\zeta\tau}{\omega} + \dfrac{1}{\omega^2} = \dfrac{\tau_1\tau_2+\tau_2\tau_3+\tau_3\tau_1}{K+1} \qquad (24)$$

$$\dfrac{\tau}{\omega^2} = \dfrac{\tau_1\tau_2\tau_3}{K+1} \qquad (25)$$

Now τ_4 appears only in equation 23. Therefore, substituting for τ into equation 24 from equation 25 and differentiating equation 24 with respect to ω, the condition for maximum ζ is obtained by equating $(d\zeta)/(d\omega) = 0$.

The following relations hold for ζ_{max}

$$\omega^2\Big|_{\zeta_{max}} = \dfrac{3(K+1)\omega_1\omega_2\omega_3}{\omega_1+\omega_2+\omega_3} \qquad (26)$$

$$\zeta_{max} = \dfrac{\omega_1+\omega_2+\omega_3}{3\omega} \qquad (27)$$

Substituting these values into equation 23

$$\tau_4\Big|_{\zeta_{max}} = \dfrac{1}{K}\left(\dfrac{3(K+1)}{\omega_1+\omega_2+\omega_3} + \dfrac{2}{3}\dfrac{(K+1)(\omega_1+\omega_2+\omega_3)}{\omega^2} - (\tau_1+\tau_2+\tau_3)\right) \qquad (28)$$

Thus the maximum damping which can be achieved by including a lead term is determined and also the value of τ_4 for which this condition holds. In practice it may sometimes be desirable to design for somewhat less than maximum damping, and derive advantage from the increased ω which results from an increased τ_4. Thus it is best to determine the maximum damping and, if it is large enough, to increase τ_4 beyond that required for maximum damping until the damping is reduced to the smallest allowable value.

This is best done by taking a larger value for ω than $\omega|_{\zeta_{max}}$ determining its corresponding τ from equation 25 and finding ζ from equation 24. τ_4 is then given by equation 23.

It is interesting to note that $\omega|_{\zeta_{max}}$ as given by equation 26 is $\sqrt{3}$ times the value of ω for the uncompensated 3-lag system, as given by the approximate equation 12. Thus, at maximum damping, the lead has also increased the undamped natural frequency of the system by a factor of approximately $\sqrt{3}$. This gives a basis for evaluating the advantages to be gained by lead compensation.

More than Three Lags in System

The same correction equation has proved to be applicable, approximately, to allow for more than three lags in the system. Thus if ζ_{III} is known, taking the three longest lag time constants and the lead ζ_{IV} is given by equation 22. Similar procedure is adopted for more than four lags.

Lead-Lag Networks

Lead-lag networks may be designed from the foregoing relations. The lag of the compensator is treated as a system lag. The design of lag-lead or lag compensators is done in a similar way.

Ratio of Pure Exponential and Oscillatory Decay Term in Transient

$$\tau\big|_{\zeta_{max}} = \frac{3}{\omega_1+\omega_2+\omega_3} \tag{29}$$

Since ω is increased and τ is increased compared to the 3-lag no-lead system, $\tau\omega$ is considerably increased. This means that the pure exponential term is not necessarily negligible any longer. The ratio of their amounts is given by

$$\rho = \frac{\tau\omega^2(\tau_4-\tau)\sqrt{(1-\zeta^2)}}{[(1-2\tau\zeta\omega+\tau^2\omega^2)(1-2\tau_4\zeta\omega+\tau_4^2\omega^2)]^{1/2}} \tag{30}$$

Usually, $\tau\omega$ is of the order of magnitude of 1 and the behavior of the system is still largely influenced by the oscillatory term. The presence of a pure exponential term with a positive coefficient which occurs if $\tau_4 > \tau_1$ increases the overshoot compared to a 3-lag system of similar damping factor. Open-loop transfer functions with more than one lead are not discussed herein. Conversely, if $\tau_4 < \tau_1$, the overshoot is reduced.

Simple Multiloop Systems with Rate Feedback

Rate feedback compensation is frequently used to improve system performance. Such rate feedback may be represented by the transfer function $(H\tau s)/(1+\tau s)$. If, however, the rate signal has a negligible delay, the transfer function can be written as $H's$.

2-Lag System with Rate Feedback

Consider a system of open-loop transfer function $K/[(1+\tau_1 s)(1+\tau_2 s)]$ and rate feedback $H(s) = (H\tau_3 s)/(1+\tau_3 s)$. The closed-loop transfer function

$$M(s) = \frac{G(s)}{1+G(s)[1+H(s)]}$$

$$= \frac{K}{K+1}\cdot\frac{1+\tau_3 s}{1+\dfrac{\tau_1+\tau_2+\tau_3[1+K(1+H)]}{K+1}s+}$$

$$\frac{\tau_1\tau_2+\tau_2\tau_3+\tau_3\tau_1}{K+1}s^2+\frac{\tau_1\tau_2\tau_3}{K+1}s^3$$

giving the equations, as before

$$\tau+\frac{2\zeta}{\omega} = \frac{\tau_1+\tau_2+\tau_3[1+K(1+H)]}{K+1} \tag{31}$$

$$\frac{2\zeta}{\omega}+\frac{1}{\omega^2} = \frac{\tau_1\tau_2+\tau_2\tau_3+\tau_3\tau_1}{K+1} \tag{32}$$

$$\frac{\tau}{\omega^2} = \frac{\tau_1\tau_2\tau_3}{K+1} \tag{33}$$

As first approximation, if τ_3 is not small

$$\tau = \tau_3\frac{1+K+KH}{K+1}\simeq(1+H) \tag{34}$$

Then from equation 33

$$\omega^2 = \frac{1+K+KH}{\tau_1\tau_2} \tag{35}$$

Substituting into equation 32 for ω^2 and τ

$$\zeta = \frac{1}{2}\omega\left(\frac{\tau_1+\tau_2}{1+K+KH}+\frac{\tau_1\tau_2}{\tau_3}\frac{KH}{(1+K+KH)^2}\right) \tag{36}$$

τ cannot be neglected now and the relative amounts of the two components of the transient need to be evaluated.

For a step displacement input, this ratio

$$\frac{A}{B} = \rho$$

$$= \frac{(\tau_3-\tau)\omega^2\tau\sqrt{(1-\zeta^2)}}{[(1-2\tau\zeta\omega+\tau^2\omega^2)(1-2\tau_3\zeta\omega+\tau_3^2\omega^2)]^{1/2}} \tag{37}$$

Thus, for large H, ω is large and ζ is small but $\rho\simeq-\tau/\tau_3\simeq(1+KH+K)/1+K$ $\simeq -H$.

Therefore, the oscillatory term is present only in small amplitudes, of high frequency, and low damping factor at the beginning of the transient. The main transient is given by τ.

With only two time constants in the forward path, this fast initial oscillation is never unstable since ζ is always positive. However, if more than two time constants are present in the forward loop and H is increased excessively, a very fast unstable oscillation occurs superimposed on the transient. Since initially this oscillation is present only in minute amplitude, it will take a short time to increase enough to be noticeable. The transient will start apparently smoothly and break into rapid, small, increasing oscillations at some point.

The correction of ζ for the presence of additional time lags in the forward path may be made by the general equation 22. Such large values of H as described in the previous paragraph are not frequently used in practice.

For values of τ_3 less than $1/\omega$ the system may be considered as a modified 3-lag system. In this case the term $(2\tau\zeta)/\omega$ is neglected as in the analysis for the 3-lag system, and correction is then
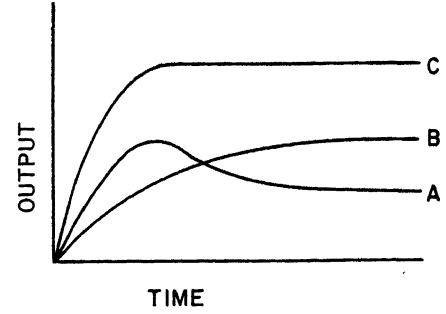
applied. For τ_3 larger than $1/\omega$, this correction method does not converge, and the foregoing analysis is applied.

Thus, for $\tau_3 < 1/\omega$

$$n = \frac{\{\tau_1+\tau_2+\tau_3[1+K(1+H)]\}(\omega_1+\omega_2+\omega_3)}{K+1} \tag{38}$$

$$\omega^2 = \frac{(K+1)\omega_1\omega_2\omega_3}{\omega_1+\omega_2+\omega_3}$$

$$\zeta = \frac{n-1}{2}\frac{\omega}{\omega_1+\omega_2+\omega_3}$$

3-Lag System with Rate Feedback

Consider the rate feedback to have no delay, so that $H(s) = Hs$. Then the closed-loop transfer function

$$M(s) = \frac{K}{K+1}\cdot\frac{1}{1+\dfrac{\tau_1+\tau_2+\tau_3+KH}{K+1}s+}$$

$$\frac{\tau_1\tau_2+\tau_2\tau_3+\tau_3\tau_1}{K+1}s^2+\frac{\tau_1\tau_2\tau_3}{K+1}s^3$$

Thus the effect of the rate feedback on the characteristic equation is the same as that of a lead of time constant H. However, the ratio ρ of the two transient components is different from the lead case, since the numerator of the closed-loop transfer function is now 1.

The equations derived for optimizing the lead time constant can be applied here; thus

$$\omega^2\big|_{\zeta_{max}} = \frac{3(K+1)\omega_1\omega_2\omega_3}{\omega_1+\omega_2+\omega_3}$$

$$= \frac{\omega_1+\omega_2+\omega_3}{3\omega}$$

for

$$H = \frac{1}{K}\left(\frac{3(K+1)}{\omega_1+\omega_2+\omega_3}+\frac{2}{3}(K+1)\times\right.$$

$$\left.\frac{\omega_1+\omega_2+\omega_3}{\omega^2}-(\tau_1+\tau_2+\tau_3)\right) \tag{39}$$

$$\rho = \frac{-\tau^2\omega^2\sqrt{(1-\zeta^2)}}{(1-2\tau\zeta\omega+\tau^2\omega^2)^{1/2}} \tag{40}$$

The optimum design is arrived at by a

judicious choice of ζ_{max}. If $\zeta_{max} \simeq 0.5$ and $\rho \simeq 1.5$, the transient for step displacement input is composed of two components that add (curves A and B, Fig. 1) to give a better transient (curve C, Fig. 1) then either component alone. It is thus possible to design the shape of the transient for these systems through consideration of the several components of the transient.

It is seen that rate feedback compensation increases the undamped natural frequency by approximately $\sqrt{3}$ for ζ_{max}. However, there results a larger coefficient of $\epsilon^{-t/\tau}$ than for lead compensation. This may be beneficial or harmful according to the adjustment of the system, and the type of inputs introduced.

ADDITIONAL LAGS IN THE SYSTEM

Additional smaller lags either in the forward path or the feedback path can be allowed for by correction of equation 22. The case of a large lag in the rate feedback has not been discussed for a forward 3-lag system.

Conclusion

A simple and powerful method is outlined here by means of which feedback systems of certain types may be designed and analyzed without recourse to either the complete Laplace method or graphical methods.

The new analytic method has the advantage of giving an over-all picture of the behavior of the system in one set of equations. A large family of solutions on an analogue computer would be required to give an equivalent picture. The effect of changes of the open-loop parameters are readily seen, and, in addition, some new insights are gained into the behavior of linear feedback systems as well.

The new approach makes direct design for transient response possible. Thus, it has an advantage over indirect frequency-response methods. In view of the ease and rapidity of its application it should be a useful addition to the present methods of design.

References

1. TRANSIENTS IN LINEAR SYSTEMS, M. F. Gardner, J. L. Barnes. John Wiley and Sons, Inc., New York, N. Y., 1942.

2. REGENERATION THEORY, H. Nyquist. *Bell System Technical Journal*, New York, N. Y., vol. 11, 1932, pp. 126–47.

3. NETWORK ANALYSIS AND FEEDBACK AMPLIFIER DESIGN, H. W. Bode. D. Van Nostrand Company, Inc., New York, N. Y., 1948.

4. CONTROL SYSTEM SYNTHESIS BY ROOT-LOCUS METHOD, W. R. Evans. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 66–69.

5. TRANSIENT RESPONSE OF MULTISTAGE VIDEO-FREQUENCY AMPLIFIERS, A. V. Bedford, G. L. Fredendall. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 27, 1939, p. 277.

6. D. C. MACHINES FOR CONTROL SYSTEMS, A. Tustin. The Macmillan Company, New York, N. Y., 1952.

7. A SERIES METHOD OF CALCULATING SERVO-MECHANISM TRANSIENT RESPONSE FROM THE FREQUENCY RESPONSE, D. V. Stallard. *M.I.T. Servomechanisms Laboratory Report No. 7138-R-2*, Massachusetts Institute of Technology, Cambridge, Mass., Nov. 16, 1953.

8. SERVOMECHANISMS AND REGULATING SYSTEM DESIGN, H. Chestnut, R. W. Mayer. John Wiley and Sons, Inc., New York, N. Y., 1951.

9. PRINCIPLES OF SERVOMECHANISMS, G. S. Brown, D. P. Campbell. Ibid., 1948, chap. 11.

10. AUTOMATIC FEEDBACK CONTROL SYSTEMS SYNTHESIS, J. G. Truxal. McGraw-Hill Book Company, Inc., New York, N. Y., 1955.

♦

Discussion

Gordon R. Slemon and J. M. Ham (University of Toronto, Toronto, Ont., Canada):
Mr. Clynes is to be complimented on the simple analytical method he has developed as an alternative to the frequency-response methods most often used for analysis and synthesis. He has presented exact expressions for the response of a number of the simpler systems and approximate expressions for the more complex types. Experience has shown that the response of many complex control systems met in practice can be approximated by a dominant damped oscillatory term and a dominant pure exponential term. There are however a number of systems for which this approximation does not hold. Further, the author's statement that "with single-loop multilag systems, more than one pair of complex roots does not arise regardless of how many lags there are in the system" is false. For example consider the open-loop transfer function

$$G(S) = \frac{K}{(1+\tau_1 S)(1+\tau_2 S)(1+\tau_3 S)(1+\tau_4 S)}$$

(41)

where $\tau_1 \approx \tau_2 \approx \tau_3 \approx \tau_4$. For example, such a case may arise when dealing with cascaded electric machines. As K is increased from zero, the roots of the closed-loop transfer function $M(S)$ come together in pairs and split off the real axis to form two pairs of conjugate complex roots having different complex frequencies. The transient response may then be characterized by a beat between the two complex frequencies and may not be accurately described by an equivalent dominant pair of roots.

In view of examples such as these, it is felt that the value of the paper would be much enhanced by a careful statement of the conditions under which his simplified analysis is not valid.

M. E. Clynes: Mr. Slemon and Mr. Ham have presented an interesting example, which can be shown to fall within the scope of the new method without difficulty. It was stated that the accuracy of the method increases with the spread of the time constants. The present example exhibits the extreme opposite of this condition, namely, four equal time constants. Nevertheless, it can be shown that the method gives good results in this case too.

The system of open-loop transfer function

$$G = \frac{K}{(1+s)(1+s)(1+s)(1+s)}$$

is absolutely unstable for $K=4$. (This, incidentally, indicates that such a system has comparatively little practical significance.)

In this case, the characteristic equation is

$$s^4 + 4s^3 + 6s^2 + 4s + 5 = 0$$

which can be written as

$$(s^2+1)(s^2+4s+5) = 0$$

The roots are thus

$\pm j$, and $-2 \pm j$

corresponding to

$\omega_1 = 1$
$\zeta_1 = 0$

and

$\omega_2 = \sqrt{5}$

$\zeta_2 = \frac{2}{\sqrt{5}} = 0.895$

where ω_1 and ζ_1, ω_2 and ζ_2 represent the undamped natural frequency and damping factor of the two pairs of roots respectively.

For $K=1$, the characteristics equation becomes

$$s^4 + 4s^3 + 6s^2 + 4s + 2 = 0$$

or

$$[s^2+(2+\sqrt{2})s+2+\sqrt{2}]$$
$$[s^2+(2-\sqrt{2})s+2-\sqrt{2}] = 0$$

Thus, in this case:

$\omega_1 = \sqrt{2-\sqrt{2}} = 0.7654$

$\zeta_1 = 1/2\sqrt{2-\sqrt{2}} = 0.3827$

$\omega_2 = \sqrt{2+\sqrt{2}} = 1.848$

$\zeta_2 = 1/2\sqrt{2-\sqrt{2}} = 0.924$

The multipliers of the transient components for a step function displacement input are

$A_1 = 1.31$
$A_2 = 0.54$

The transient components and the total transient are shown in Fig. 2.

As the gain is increased from 1 to the point of instability, 4, ζ_2 changes from 0.924 to 0.895.

It can be seen, therefore, that the discussers are not correct when they state that the transient can be regarded as a beat between the two frequencies. Damping
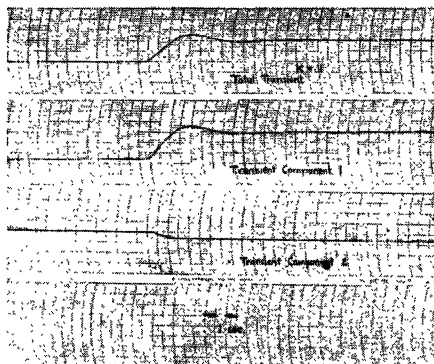
Fig. 2. Transient components and total transient for a step function displacement input for the system

$$G = \frac{K}{(1+s)(1+s)(1+s)(1+s)} \text{ for } K=1$$

Fig. 3. Transient response for a step function displacement input for the system

$$G = \frac{K}{(1+s)(1+s)(1+s)(1+s)} \text{ for } K=1, 2, \text{ and }$$

4.

Fig. 4. Transient response for a step function displacement input for the system

$$G = \frac{K}{(1+0.1s)(1+0.1s)(1+0.1s)(1+0.1s)(1+0.1s)(1+0.1s)(1+0.1s)}$$

factors of the order of 0.9 are so close to critical damping, that, when they occur in minor transient components, they have a quite negligible effect on the degree of stability and on the type of transient response.

There is no doubt that, for this system, the transient is characterized chiefly by the dominant pair of roots.

Fig. 3 shows the transient response for a step function displacement for values of K of 1, 2, and 4. In no case is any noticeable error committed by regarding the secondary roots as pure exponentials.

The statement queried by Mr. Slemon and Mr. Ham is correct in the sense it was meant to convey namely that for single-loop multilag systems more than one pair of significant oscillatory roots do not arise.

To apply the new method to this system, the system is first treated as a 3-lag system. Taking the case of $K=1$,

$$\omega_{III}^2 = \frac{2}{3}, \quad n_{III} = 4.5$$

$\zeta_{III} = 0.476$
giving a corrected value of

$$\omega_{III}'^2 = \frac{\frac{2}{3}}{1 - \frac{4 \times 0.476^2}{3.5}} = \frac{\frac{2}{3}}{0.743} = 0.897$$

giving

$$\omega_{IV}^2 = 0.897 \times 1/2 = 0.449$$

$\omega_{IV} = 0.671$

For ζ, use the corrected value of ζ_{III} which is 0.501 and then

$\zeta_{IV} = 0.501(1 - 0.25)$
$= 0.376$

This compares with the actual values of $\omega = 0.765$ and $\zeta = 0.383$.

It is seen that the new method gives a somewhat lower value for ω but gives a close approximation for ζ, erring in each case on the side of safety. It is to be emphasized that the present example represents the least favorable condition for accuracy.

The method also gives an indication of the magnitude of minor components of the transient.

The corrected value for τ_{III} is 0.457 corresponding to an ω of 2.23. This is a rough guide to the value of ω_2 which is 1.85. A better approximation is arrived at by reducing the value 2.23 by the same ratio as was used to evaluate ω_{IV} from ω_{III}. The resulting value for ω_2 is 1.76. However, the general applicability of such a procedure needs to be investigated further.

As an indication of the usefulness of the method for high order systems, consider for example a system such as the discussers suggest, but with seven time constants such that

$$\frac{K}{(1+\tau_1 s)(1+\tau_2 s)(1+\tau_3 s)\dots(1+\tau_7 s)}$$

$\tau_1 = \tau_2 = \tau_3 = \tau_4 = \dots = \tau_7 = 0.1$ sec. For $K = 1$,
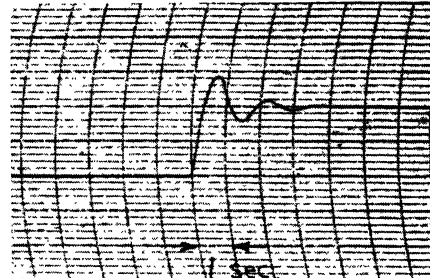
$$\omega_{VII} = \frac{2}{0.21}$$

Corrected value

$$\omega_{VII}'^2 = \frac{2}{0.21} \times \frac{1}{0.743}$$

or

$\omega_{VII} = 3.58$ giving a period $\frac{2\pi}{\omega_{VII}}$ of 1.75 sec.

$$\zeta_{OII} = \zeta_{III}\left(1 - \frac{4K}{7}\right) = 0.501 \times \frac{3}{7}$$
$$= 0.215$$

The computer result for the same system is shown in Fig. 4 and it can be seen that it closely agrees with the calculated values. It will be appreciated that this example can be solved only with great laboriousness by any other known method, graphical or analytical.

In conclusion, the author wishes to point out that the response of any system is the total of the individual transient components. Thus, occasionally, a system without oscillatory roots may exhibit overshoot if some of the multipliers of transient components are of opposite signs. Once the roots are known, these multipliers can always be determined from inverse Laplace transfer tables and the total transient constructed from the individual components.

The author does not know of any exceptions for which the method given is not valid. He intends to elucidate and amplify the method further in a subsequent paper.

An Analysis and Analogue-Computer Study of a Force-Reflecting Positional Servomechanism

M. G. SPOONER
ASSOCIATE MEMBER AIEE

C. H. WEAVER
MEMBER AIEE

THE problem of force reflection, or feel, has been of interest to designers of remotely controlled manipulators for the past several years. Systems have been designed and built which demonstrate both the possibilities and difficulties inherent in any attempt to obtain such feel characteristics.[1,2] In addition, more complicated systems have been built in attempts to overcome such difficulties,[3] and reports outlining the general feel problem of manipulators and suggest courses of further study have appeared in the literature.[4] It is the purpose of this paper to present an analogue study of a basic force-reflecting positional system along with a companion dynamic analysis upon which any desired stability or synthesis study may be based.

Analysis and Synthesis Discussion

The system to be considered herein is shown in semischematic form in Fig. 1 along with the system variables and their positively assumed directions. The meanings of the various symbols are apparent from their positions on the diagram The system consists of two conventional positional servomechanisms, so connected that the output shaft of servomechanism 1 is the input shaft of servomechanism 2, and vice versa. Thus, for a specified input signal at either shaft, the system attempts to reproduce that signal at the other shaft in the usual servomechanism manner. If such reproduction is prevented because of a force applied at the output shaft, a similar force then appears at the input shaft, related to the preventional force in a definite way. This is essentially the feel characteristic of the system.

The interactions existing between such

forces as well as various other pertinent relations between the variables shown in Fig. 1 are given in the following equations. The $KG(s)$ notation is used since the analysis is facilitated by separating gain from dynamics.

The system equations in the time domain are

$$J_1\ddot{\theta}_1(t)+f_1\dot{\theta}_1(t)=K_1\epsilon_1(t)+T_1(t) \quad (1)$$

$$J_2\ddot{\theta}_2(t)+f_2\dot{\theta}_2(t)=K_2\epsilon_2(t)+T_2(t) \quad (2)$$

Taking the Laplace transform and substituting for $\epsilon_1(t)$ and $\epsilon_2(t)$ in terms of $\theta_1(t)$ and $\theta_2(t)$, the system equations in the s domain are

$$J_1s^2\theta_1(s)-J_1s\theta_1(0+)-J_1\dot{\theta}_1(0+)+f_1s\theta_1(s)-f_1\theta_1(0+)=K_1[\theta_2(s)-\theta_1(s)]+T_1(s) \quad (3)$$

$$J_2s^2\theta_2(s)-J_2s\theta_2(0+)-J_2\dot{\theta}_2(0+)+f_2s\theta_2(s)-f_2\theta_2(0+)=K_2[\theta_1(s)-\theta_2(s)]+T_2(s) \quad (4)$$

Let

$$G_1(s)=\frac{1}{J_1s^2+f_1s}$$

and

$$G_2(s)=\frac{1}{J_2s^2+f_2s}$$

Using these definitions and after some manipulation, equation 3 becomes

$$[1+K_1G_1(s)]\theta_1(s)-K_1G_1(s)\theta_2(s)=G_1(s)T_1(s)+J_1G_1(s)\dot{\theta}_1(0+)+G_1(s)(J_1s+f_1)\theta_1(0+) \quad (5)$$

In similar fashion, equation 4 becomes

$$-K_2G_2(s)\theta_1(s)+[1+K_2G_2(s)]\theta_2(s)=G_2(s)T_2(s)+J_2G_2(s)\dot{\theta}_2(0+)+G_2(s)(J_2s+f_2)\theta_2(0+) \quad (6)$$

Solving equations 5 and 6 by determinants for the two unknowns, $\theta_1(s)$ and $\theta_2(s)$, gives the two following equations

$$\theta_1(s)=\frac{[1+K_2G_2(s)]G_1(s)T_1(s)}{1+K_1G_1(s)+K_2G_2(s)}+$$
$$\frac{J_1[1+K_2G_2(s)]G_1(s)\dot{\theta}_1(0+)}{1+K_1G_1(s)+K_2G_2(s)}+$$
$$\frac{[1+K_2G_2(s)]G_1(s)(J_1s+f_1)\theta_1(0+)}{1+K_1G_1(s)+K_2G_2(s)}+$$
$$\frac{K_1G_1(s)G_2(s)T_2(s)}{1+K_1G_1(s)+K_2G_2(s)}+$$

$$\frac{J_2K_1G_1(s)G_2(s)\dot{\theta}_2(0+)}{1+K_1G_1(s)+K_2G_2(s)}+$$
$$\frac{K_1G_1(s)G_2(s)(J_2s+f_2)\theta_2(0+)}{1+K_1G_1(s)+K_2G_2(s)} \quad (7)$$

$$\theta_2(s)=\frac{[1+K_1G_1(s)]G_2(s)T_2(s)}{1+K_1G_1(s)+K_2G_2(s)}+$$
$$\frac{J_2[1+K_1G_1(s)]G_2(s)\dot{\theta}_2(0+)}{1+K_1G_1(s)+K_2G_2(s)}+$$
$$\frac{[1+K_1G_1(s)]G_2(s)(J_2s+f_2)\theta_2(0+)}{1+K_1G_1(s)+K_2G_2(s)}+$$
$$\frac{K_2G_2(s)G_1(s)T_1(s)}{1+K_1G_1(s)+K_2G_2(s)}+$$
$$\frac{J_1K_2G_2(s)G_1(s)\dot{\theta}_1(0+)}{1+K_1G_1(s)+K_2G_2(s)}+$$
$$\frac{K_2G_2(s)G_1(s)(J_1s+f_1)\theta_1(0+)}{1+K_1G_1(s)+K_2G_2(s)} \quad (8)$$

Equations 7 and 8 are in a form readily applicable to both analysis and synthesis. Complete solutions for any combination of input and boundary condition forcing can be directly calculated. The system characteristic equation is $1+K_1G_1(s)+K_2G_2(s)=0$, and, thus, synthesis can be carried out in the usual manner with a plot of $K_1G_1(s)+K_2G_2(s)$. Since sums instead of products are involved, polar plots would appear most useful. Compensation networks can be added in either or both channels, and ω_r and gain and phase margin can be given the usual interpretation.

An interesting situation arises when the two channels are identical. In such case, $K_1G_1(s)=K_2G_2(s)$, and the system characteristic equation becomes $1+2K_1G_1(s)=1+2K_2G_2(s)=0$. Thus, if the system is composed of two identical servomechanisms synthesized separately, it would be less stable than one of the servomechanisms operated alone, by a factor of two in gain.

Precisely this effect is described in reference 1 where such a design technique was used, and it was found that the system with both outputs free was considerably less stable than with one output fixed. Verification of this effect is also shown in case A in the following, in result 1, and in Fig. 2.

Analysis of Special Cases

The following special cases are of interest both to the previous analysis and synthesis discussion and to the subsequent analogue study.

CASE A

In this case the assumption is made that $\dot{\theta}_1(0+)=0$, $\dot{\theta}_2(0+)=0$, $T_1(t)=0$, $T_2(t)=0$, $\theta_2(0+)=\theta_2$, and $\theta_1(0+)=0$. Using these relations, obtains the following equations
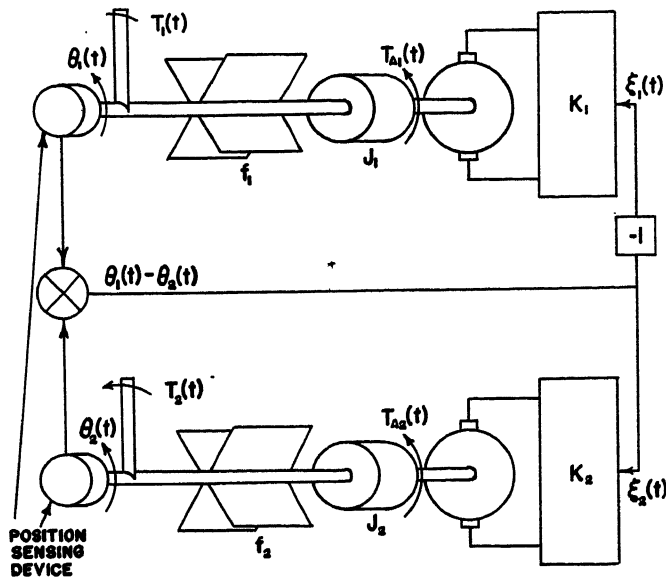
Fig. 1 (left). System block diagram of force-reflecting positional servomechanism

Fig. 3 (right). Analogue computer block diagram

A—Adder
C—Scale changer
J—Integrator
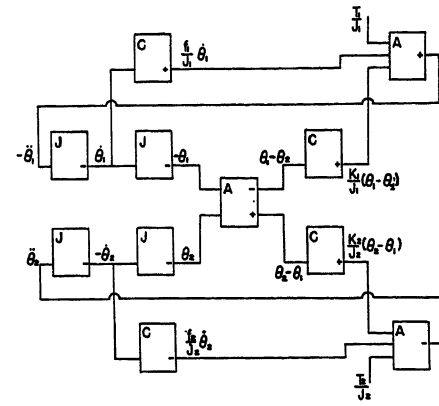
$$\theta_1(s) = \frac{K_1 G_1(s) G_2(s)(J_2 s + f_2)\theta_2}{1 + K_1 G_1(s) + K_2 G_2(s)} \quad (9)$$

$$\theta_1(s) = \frac{K_1 G_1(s)}{1 + K_1 G_1(s) + K_2 G_2(s)} \frac{\theta_2}{s} \quad (10)$$

Thus, for an initial displacement of the two shafts and following a release at $t=0$, the system responds as though it were a single loop servo with a step input θ_2 at $t=0$, with forward and feedback blocks obtainable from equation 10. In particular, if $K_1 G_1(s) = K_2 G_2(s)$, the forward block is $K_1 G_1(s) = K_2 G_2(s)$, and the feedback block is two. The fact that the initial condition appears as a step input is an interesting application of the material presented in reference 5.

Fig. 2. Comparison of stability of double- and single-loop cases

A—$\theta_2(t)$ response for double-loop case with an initial displacement of $\theta_1(t)$ at t=0
B—$\theta_2(t)$ response for single-loop case with a step input of position at $\theta_1(t)$
C—Same situation as (B), except that K_2 is doubled

Applying the final value theorem to equation 10, the result is obtained that

$$\theta_1(t)_{ss} = \frac{\theta_2 K_1 f_2}{K_1 f_2 + K_2 f_1} \quad (11)$$

in particular, if $K_1 G_1(s) = K_2 G_2(s)$

$$\theta_1(t)_{ss} = \frac{\theta_2}{2}$$

a result to be expected from a physical concept.

CASE B

In this case the assumption is made that $T_1(s) = T_1/s$, $T_2(s) = 0$, $\dot\theta_2(0+) = 0$, $\dot\theta_1(0+) = 0$, $\theta_2(0+) = 0$, and $\theta_1(0+) = 0$. With the use of these relations, the following equations are obtained

$$\theta_1(s) = \frac{[1 + K_2 G_2(s)]}{1 + K_1 G_1(s) + K_2 G_2(s)} G_1(s) \frac{T_1}{s} \quad (12)$$

with the final value theorem $\lim \theta_1(t) = \infty$. In a similar fashion as $t \to \infty$ for the same case, $\lim \theta_2(t) = \infty$.

Thus a step function of torque applied to one of the shafts (or the existence of a steady difference in torques applied to the two shafts) results in unlimited motion at each shaft.

Further equation manipulation shows that

$$e_1(s) = \theta_2(s) - \theta_1(s) = -\frac{G_1(s)T_1(s)}{1 + K_1 G_1(s) + K_2 G_2(s)} \quad (13)$$

With the step input of torque as before,

$$e_1(t)_{ss} = \frac{T_1 f_2}{K_1 f_2 + K_2 f_1} \quad (14)$$

Thus the error assumes a constant value in the steady state when a step input of torque is applied.

CASE C

Suppose that an impulse of torque is applied to the system, the conditions being the same as in case B. Then $T_1(s) = T_1$, and by using the final value theorem on the appropriate equations

$$\theta_1(t)_{ss} = \frac{K_2 T_1}{K_1 f_2 + K_2 f_1} \quad (15)$$

$$\theta_2(t)_{ss} = \frac{K_2 T_1}{K_1 f_2 + K_2 f_1} \quad (16)$$

$$\epsilon(t)_{ss} = 0 \quad (17)$$

Analogue Computer Study

The analogue computer study is based on equations 1 and 2. When they are normalized for computer convenience:

$$\ddot\theta_1(t) = \frac{T_1(t)}{J_1} - \frac{K_1}{J_1} [\theta_1(t) - \theta_2(t)] - \frac{f_1}{J_1} \dot\theta_1(t) \quad (18)$$

$$\ddot\theta_2(t) = \frac{T_2(t)}{J_2} - \frac{K_2}{J_2} [\theta_2(t) - \theta_1(t)] - \frac{f_2}{J_2} \dot\theta_2(t) \quad (19)$$

Fig. 4. Results of a step function of torque applied to the $\theta_1(t)$ shaft

A—Step function of torque applied to the $\theta_1(t)$ shaft
B—$\theta_1(t)$ response to the input of (A)
C—$\theta_2(t)$ response to the input of (A)
D—Error $K_1/J_1[\theta_1(t) - \theta_2(t)]$ caused by the input of (A)
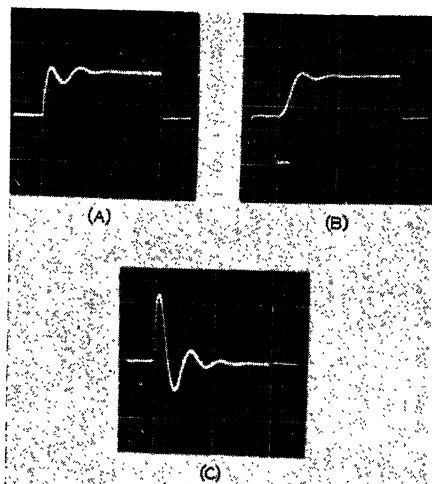
Fig. 5. Results of impulse function of torque applied to the $\theta_1(t)$ shaft

A—$\theta_1(t)$ response to the torque impulse input
B—$\theta_2(t)$ response to the torque impulse input
C—Error $K_1/J_1[\theta_1(t)-\theta_2(t)]$ caused by the torque impulse input

The analogue computer connection diagram derived from these equations and used to study the system is shown in Fig. 3. The results obtained from this study are discussed in the following.

RESULT 1

In case A, discussed in the section entitled "Analysis and Synthesis Discussion," a factor of two in gain existed between the characteristic equation of a single-loop servomechanism and the characteristic equation of a feel system composed of two such servomechanisms. To check this on the analogue computer, the double-loop case was studied by putting an initial displacement on $\theta_1(t)$ at $t=0$, then observing $\theta_2(t)$ for $t>0$. The result is shown in Fig. 2. Then the single-loop case was studied by opening the no. 1 feedback, and putting a step displacement on $\theta_1(t)$ equal to the initial displacement previously used. The value of $\theta_2(t)$ for this case is also shown in Fig. 2. These curves indicate clearly that the single-loop case is much more stable. As a last check, the gain of the no. 2 loop was doubled, and the response again observed. This is shown in Fig. 2 and checks closely with the result in the double-loop case shown which is in the same figure.

RESULT 2

Result 2 is the analogue study corresponding to case B. A step function of torque was applied to the θ_1 shaft, and $\theta_1(t)$, $\theta_2(t)$, and $\epsilon_1(t)$ were observed. The results are shown in Fig. 4 and they check the calculated values.

RESULT 3

Case C analyzed the system with an impulse of torque applied to the θ_1 shaft. This situation was duplicated on the computer and the values of $\theta_1(t)$ and $\theta_2(t)$ corresponding to such an input are shown on Fig. 5. The displayed values check the analysis.

RESULT 4

An interesting case concerns the system response when, with a step input of torque, the output member strikes a fixed object. This situation can be handled nicely by the use of a Philbrick bounding unit.[6] The torque input was applied to the θ_1 shaft and the bounding unit was used to limit the motion of θ_2 to some fixed maximum value. Fig. 6 shows the results, the point of impact being clearly indicated by the breaks in the $\theta_1(t)$, $\theta_2(t)$, and $\epsilon(t)$ curves.

RESULT 5

The results of this portion of the analogue computer study will be of greater use if a discussion of force reflection or feel is first included.

Discussion of Feel

Although the idea of force reflection, or feel, is implied throughout the previous work, only result 4 actually gives any indication of such a characteristic. Feel is demonstrated in this result by the input-position limiting as a result of output-position limiting. In Fig. 6, concerned with this result, the curves imply that before the output member strikes the fixed object, all the applied torque is error torque, i.e., since no torque is present at the output, any torque present at the input is in error from the desired value of zero. On the other hand, after the output member strikes the fixed object and both $\theta_1(t)$ and $\theta_2(t)$ have become constant, the steady-state torque ratio can be easily calculated. Since no motion exists, the output torque is $K_2\epsilon_2$, and the input torque is $K_1\epsilon_1$. Substituting, $T_{out}=K_2(\theta_1-\theta_2)$ and $T_{applied}=K_1(\theta_2-\theta_1)$. Thus $T_{out}=(-K_2/K_1)T_{applied}$, and this relation can be used by an operator to determine the force at the output member corresponding to the force applied at the input under this condition of no motion. In particular, a 1-to-1 relation is obtained if the two are identical.

Just as the "before" case indicated complete error, so the "after" case can be considered to indicate no error provided K_2 and K_1 do not change. From these considerations it would seem feasible to define a feel error as being the dif-

ference between the applied force as a function of time, and the output member force as a function of time multiplied by the function K_2/K_1. As an equation, this would give

$$\epsilon_T(t)=T_A(t)-\frac{-K_2}{K_1}T_R(t) \tag{20}$$

where

ϵ_T = feel or torque error
T_A = applied torque
T_R = retarding torque

An interesting difficulty arises concerning the use of this equation. Thus, if two input torques are defined, the $\epsilon_T(t)$ is fixed by definition and no useful information is obtained. On the other hand, if a given angular input $\theta_1(t)$ is defined along with $T_R(t)$, the system is by definition a single-loop system. However, the force felt by the operator producing the specified $\theta_1(t)$ can be calculated as follows

$$T_A(t)=K_1[\theta_1(t)-\theta_2(t)]-J_1\ddot{\theta}_1-f_1\dot{\theta}_1 \tag{21}$$

The system equations can be solved for $\theta_2(t)$ with the given inputs $\theta_1(t)$ and $T_R(t)$, and the $\theta_2(t)$ value can be substituted in the foregoing equation to give $T_A(t)$; or the equations can be solved directly for $T_A(t)$. In either case, the $\epsilon_T(t)$ can be computed for all values of time and studies of the feel dynamics of a given system can be based on the results.

The problem of a specified angular
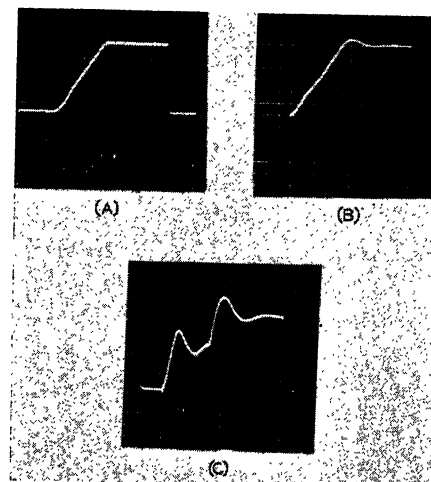
Fig. 6. Results of a step function of torque applied to the $\theta_1(t)$ shaft with $\theta_2(t)$ motion limited

A—$\theta_2(t)$ response to the torque input. The point at which limiting occurs is shown by the break to horizontal in the curve
B—$\theta_1(t)$ response to the torque input. The point at which $\theta_2(t)$ strikes its limit is shown by the flattening of the curve
C—Error, $K_1/J_1[\theta_1(t)-\theta_1(t)]$ caused by the torque input. The point at which $\theta_2(t)$ strikes its limit is shown by the sharp upward break of the curve

Spooner, Weaver—Study of a Force-Reflecting Positional Servo
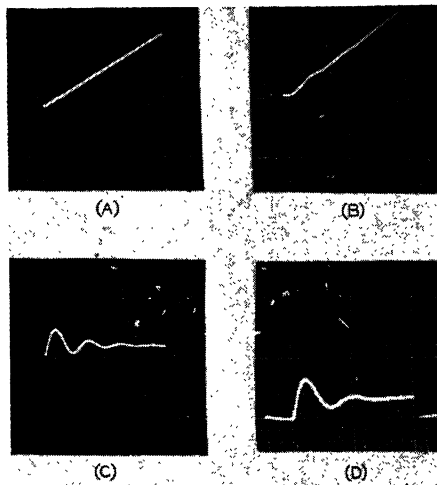
Fig. 7. Results of a ramp function of $\theta_1(t)$ applied to the $\theta_1(t)$ shaft

A—Ramp function of position applied to the $\theta_1(t)$ shaft

B—$\theta_1(t)$ response to a ramp applied to the $\theta_1(t)$ shaft

C—Error $K_1/J_1[\theta_1(t) - \theta_2(t)]$ caused by a ramp function applied to $\theta_1(t)$ shaft

D—Torque felt by the agent applying the ramp function to $\theta_1(t)$ shaft. Feel torque is $K_1/J_1[\theta_1(t) - \theta_2(t)] - (f_1/J_1)\dot\theta_1(t) - \ddot\theta_1(t)$

input to θ_1 is readily studied on the analogue computer. If a ramp function of position is used, then θ_2 will attempt to follow θ_1 with some constant error. The torque that the operator must apply at the input is given by equation 21. These results are shown in Fig. 7. Even though no opposing torque is present at the output, the operator must apply a torque to maintain the specified motion. Thus, under these conditions, the system exhibits an error in the force reflection which is proportional to the velocity of the input member.

The analogue computer study is more advantageous when it can be extended beyond the realm of straightforward analysis. If a nonlinearity is introduced, such as the bounding of the output, then the analogue computer is an extremely useful tool. In this portion of the study as well as in result 4, the output motion was bounded and interesting results were observed which extended the results beyond any previous analysis. With a ramp function of position of input specified, the motion of the output member was limited, as shown in Fig. 8. The torque required to maintain the specified input position, which is the feel torque as given by equation 21, was observed and appears also in Fig. 8. As it would be expected, the feel torque rises rapidly at the moment the output motion stops, and the operator feels the obstruction at the output.

Conclusions

An analysis and analogue computer study of the basic positional feel system shown in Fig. 1 has been presented. In the analysis the stability problem has been resolved to a degree that a synthesis can be carried out in the usual manner. In addition, various steady-state relations of interest have been considered, static feel has been discussed, and some progress was made in the study of dynamic feel.

Both analysis and analogue studies clearly show that the system has inherent positional qualities comparable with those of its constituent servomechanisms, and that static force reflection or feel is transmitted between the two shafts with an open-loop characteristic. In addition, Figs. 6 and 8 clearly show the dynamic effects of having the output member strike a fixed object—a most realistic test.

Various interesting aspects of the problem of feel transmission have not been considered. In many practical applications, the worm gears necessary in connecting an output motor to its load would render this system useless. A scheme has been studied for by-passing this difficulty.[3] The effects of changing inertias[4] and loads on the end of the output arm which make the load torque a function of the shaft angle also have not been considered. In addition, the general problem of dynamic feel, as well as the specific problem of closing the loop around such feel with the attendant advantage of elimination of calibration, seems to offer interesting possibilities for study. Since these and other considerations appear amenable to analogue-computer study, further paths of investigation seem to be clearly defined.

References

1. MANIPULATOR COMMITTEE REPORT, F. Ring, R. Fox, W. S. Ladniak, B. W. Kinyon, C. H. Weaver. *Engineering and Maintenance Division Report*, Oak Ridge National Laboratory, Oak Ridge, Tenn., 1950.

2. SERVOS FOR REMOTE MANIPULATION, R. C. Goertz, F. Bevilacqua. *National Convention Report*, Institute of Radio Engineers, New York, N. Y., session 32, pt. 9, June 17, 1953, pp. 103–09.

3. THE USE OF RESISTANCE STRAIN GAGES IN OBTAINING ARTIFICIAL FEEL FOR A REMOTE POSITIONAL SERVOMECHANISM, Stanley Sporn. Thesis, University of Tennessee, Knoxville, Tenn., 1951.

4. REMOTE CONTROL ENGINEERING, H. L. Hull, R. C. Goertz, F. Bevilacquac. *Nucleonics*, New York, N. Y., Nov. 1952, pp. 33–44.

5. INITIAL CONDITIONS IN TRANSIENT ANALYSIS, L. Tasny-Tschiassny. *Wireless Engineer*, London, England, vol. 30, Feb. 1953.

6. CATALOG AND MANUAL. George A. Philbrick Researches, Inc., Boston, Mass., 1951.
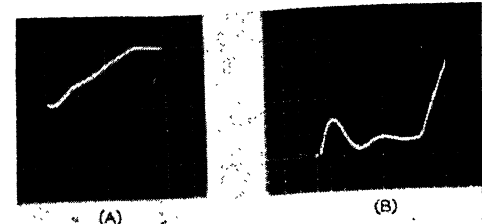
Fig. 8. Results of a ramp function of $\theta_1(t)$ applied with $\theta_2(t)$ motion limited

A—$\theta_2(t)$ response to the ramp function. The sharp break to horizontal in the curve shows where $\theta_2(t)$ strikes its limit

B—Torque felt by the agent applying the ramp function to $\theta_1(t)$ shaft. The sharp upward break shows the point at which $\theta_2(t)$ strikes its limit. Feel torque $= K_1/J_1 [\theta_1(t) - \theta_2(t)] - f_1/J_1 \dot\theta_1(t) - \ddot\theta_1(t)$

———◆———

Discussion

R. C. Goertz (Argonne National Laboratory, Lemont, Ill.): Two different and complete Master-Slave manipulators have been constructed at Argonne National Laboratory using servo mechanisms similar to the one described in the paper. The Master-Slave Servo-Manipulator, model 1, was exhibited at the Atoms for Peace Exposition, University of Michigan, Ann Arbor, Mich., in June 1954, and at the New Products, New Methods and Patents Exhibit in Detroit, Mich., November 1954. A paper describing the Master-Slave Servo-Manipulator, model 2, along with two mechanical manipulators was presented at the International Conference on the Peaceful Uses of Atomic Energy at Geneva, Switzerland, in August 1955.

Although the authors state that the system might be extrapolated to almost any size, there are certain practical limits that would limit this extrapolation when used for handling radioactive materials. This limit is placed at somewhere between 25 and 50 pounds due to the rapid increase in the effective inertia of larger motors. When these manipulators are used on a large variety of apparatus, the effective mass must be kept low enough to prevent damage when the manipulator collides with the apparatus. To overcome this limitation, we are working on special hydraulic force-reflecting servomechanisms. Under this condition we believe the extrapolation can be carried out to give a force and/or motion magnification quite possible up to 100 or so. For most of our work this ratio would generally be less than five.

M. G. Spooner and C. H. Weaver: The authors appreciate the pertinent and authoritative comments of Mr. Goertz concerning the practical uses of the system analyzed in this paper. This work was initiated by the authors because of the belief that a complete analysis of the system would serve a useful purpose in the realm of closed-loop control systems and Mr. Goertz's statements appear to justify this belief.

Estimating Transient Responses from Open-Loop Frequency Response

GEORGE A. BIERNSON
ASSOCIATE MEMBER AIEE

THE SYNTHESIS of a feedback control system is usually performed in terms of its frequency response, but system performance is best evaluated in terms of its transient response. Consequently, it is desirable to be able to relate transient response as directly as possible to frequency response. Although it is difficult to derive exact yet general mathematical relationships between these two responses, a great many reliable approximations can be found. This paper developed many of the most important of these approximations by examining and comparing the transient responses for a number of different types of feedback control transfer functions.

Relationships between the step response and frequency response of feedback control systems have been considered many times, but in spite of this there has not been presented adequate correlation between the two. The engineer generally knows that the speed of the step response is in some way related to the bandwidth of the system, but he would probably find great difficulty in defining this relationship quantitatively, except for very much idealized cases.

More important, however, than this lack of knowledge of the correlation between step response and frequency response is the more general ignorance of the relationships between the frequency response and the other transient responses of the system: the impulse response, the ramp response, the response to a step of acceleration, etc. These transient responses can usually be estimated from the frequency response even more readily than can the step response. Probably the main reason for the neglect of these transient responses is that there has not been adequate appreciation of their importance in terms of the general time behavior of feedback control systems.

Reference 1 has demonstrated the importance of the various transient responses of a feedback control system. It has shown that for a particular feedback control system there are only a few transient responses that are significant, and if all of them are known, rather rough approximations of each is usually all that

is needed for the designer to compute easily and with reasonable accuracy a plot of the system response to an arbitrary input. Consequently, if the designer can estimate these transient responses adequately by inspecting the open-loop frequency response, he is able to relate the networks he builds into a feedback control loop directly to the performance of the resultant system for any input. It is the purpose of this paper, in combination with reference 1, to achieve this end.

Since it is the open-loop characteristics that are synthesized in feedback control design, it is fortunate that most of the important characteristics of the transient responses can be estimated best by examining the open-loop frequency response. The major exception is the peak overshoot of the step response, which can be estimated best from the maximum magnitude (or M_p) of the closed-loop frequency response. In the paper, the rules for estimating these transient characteristics are developed on empirical and intuitive bases, by examining a number of different transfer functions. The author has been able to obtain a reasonable amount of theoretical justification for these rules, too lengthy to be included in the paper. The justification is based upon the material in reference 2, which shows how the closed-loop poles of a feedback control loop are related to its open-loop frequency response, and it will be presented in a future paper. Most of the material in this paper was originally presented in part IV of reference 3.

Terminology

The feedback control terms and symbols used in this paper are illustrated in Fig. 1. The block diagram represents a feedback control loop, having the variables

X_i = loop input
X_e = loop error (or loop difference)
X_u = loop return signal

The loop return signal is often the loop output X_0. The major transfer functions for the feedback-control loop are

$G = X_u/X_e$ = loop transfer function

X_u/X_i = return transfer function
X_e/X_i = error (or difference) transfer function

The loop transfer function G is often called the open-loop transfer function. The magnitude of G for real frequencies is called the loop gain.

The most important dynamic parameter of a feedback control loop is its bandwidth; and the author has found that an excellent criterion for the bandwidth of a feedback control loop is the gain crossover frequency, denoted ω_{cg}, and defined as the frequency where the loop gain is unity. The gain crossover frequency has the following advantages over other criteria for bandwidth:

1. It is easy to measure and calculate.

2. It gives a very reliable indication of the rise time of the step response for all types of feedback control systems, as is illustrated in this paper.

Another useful bandwidth parameter related to the gain crossover frequency ω_{cg} is the asymptote crossover frequency, denoted ω_c, and defined as follows:

For those feedback-control loops in which the loop gain in the region near gain crossover approaches an asymptote inversely proportional to frequency, the asymptote crossover frequency ω_c is the frequency where that asymptote is unity.

The frequencies ω_c and ω_{cg} are illustrated in Fig. 1.

In place of the decibel the new and more general logarithmic unit, the decilog, is used. For any nondimensional ratio A the value in decilogs is given by

$$A \text{ in decilogs} = 10 \log_{10} A \qquad (1)$$

Thus, a factor of 10 represents 10 decilogs and is abbreviated as 10 dlg.

Elements of Transient Response

Reference 1 has shown that there are three basic sets of characteristics which define the general time behavior of a feedback control loop; and these are the transient terms, designated $T_n(t)$, the steady-state coefficients c_n, and the initial-value coefficients a_m. These characteristics appear in the return response

Fig. 1. Illustration of terminology

FREQUENCY (LOG SCALE)
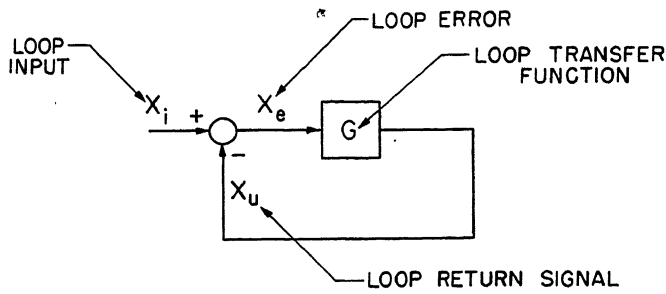
to a unit step of a derivative or an integral of the input. The return response of a feedback control loop for a unit step of the nth derivative can be expressed as follows

$$x_u(t)=T_n(t)+c_n+c_{n-1}\frac{t}{1!}+c_{n-2}\frac{t^2}{2!}+$$
$$\ldots+\ c_1\frac{t^{n-1}}{(n-1)!}+c_0\frac{t^n}{n!}\quad(2)$$

The transient portion $T_n(t)$ is called the nth-derivative return response transient term; and a coefficient c_n is called the nth-derivative return response steady-state coefficient. For a unit step of the mth integral, the return response is

$$x_u(t)=T_{-m}(t)+a_{m-1}u_1(t)+a_{m-2}u_2(t)+$$
$$\ldots+a_0u_m(t)\quad(3)$$

The symbol $u_1(t)$ represents a unit impulse at $t=0$; $u_2(t)$ represents a unit doublet; $u_3(t)$ a unit triplet, etc. A coefficient a_m is called the initial-value coefficient for the mth derivative.

The steady-state and initial-value coefficients are related to the initial values of the corresponding transient terms. If a system has a finite bandwidth, the initial values of the response to a step and the responses to steps of a derivative must be zero. Setting the value of equation 2 equal to zero at $t=0$ gives

$$c_n=-T_n(0),\text{ for }n\geq0\quad(4)$$

It can also be shown that the initial-value coefficient is equal to

$$a_m=T_{-m}(0),\text{ for }m>0\quad(5)$$

If the system has a finite bandwidth, the step response initial-value coefficient a_0 is zero. The first two terms of the expansion of equation 2 are $[T_n(t)+c_n]$, and equation 4 shows that this sum represents a response that starts at zero. This property of starting at zero is very convenient, and consequently the sum $[T_n(t)+c_n]$ is often plotted instead of the transient term $T_n(t)$ and is termed the composite transient for the nth derivative.

A step of a derivative represents a low-frequency input, and for low-frequency

inputs the return response follows the input so closely it is often desirable to consider the error response rather than the return response. The loop error is related to the loop return by

$$x_e(t)=x_i(t)-x_u(t)\quad(6)$$

Substituting the expansion of equation 2 for $x_u(t)$ into this relation, gives the following expansion for $x_e(t)$

$$x_e(t)=T_n'(t)+c_n'+c_{n-1}'\frac{t}{1!}+c_{n-2}'\frac{t^2}{2!}+$$
$$\ldots+c_1'\frac{t^{n-1}}{(n-1)!}+c_0'\frac{t^n}{n!}\quad(7)$$

where

$$T_n'(t)=-T_n(t)\quad(8)$$
$$c_n'=-c_n,\text{ for }n>0\quad(9)$$
$$c_0'=1-c_0\quad(10)$$

The transient $T_n'(t)$ is the error response transient term for the nth derivative; and a coefficient c_n' is the error response steady-state coefficient for the nth derivative, which is popularly known as the nth-derivative error coefficient.

For systems with only single-order poles, the transient terms can be computed from

$$T_n(t)=\sum_{l=1}^{L}\frac{K_{l0}}{(s_l)^n}\ e^{s_l t}\quad(11)$$

where s_l is any one of the L poles of the system transfer function; and K_{l0} represents the step response coefficient for that pole, and is equal to

$$K_{l0}=\frac{(s-s_l)}{s}\ \frac{X_u}{X_i}\bigg|_{s=s_l}\quad(12)$$

Regardless of the order of the system poles, the transient terms of successive derivatives are related by

$$T_n(t)=\int T_{n-1}(t)dt\quad(13)$$

Thus, if any transient term is specified with sufficient accuracy, all the other transient terms can be computed from it by successive differentiation and by successive integration.

The steady-state coefficients can be computed from

$$c_n=\frac{1}{n!}\frac{d^n}{ds^n}\frac{X_u}{X_i}\bigg|_{s=0}\quad(14)$$

The initial-value coefficients can be computed by replacing s in X_u/X_i by $1/\lambda$ and employing the relation

$$a_m=\frac{1}{m!}\frac{d^m}{d\lambda^m}\frac{X_u}{X_i}(1/\lambda)\bigg|_{\lambda=0}\quad(15)$$

On the other hand, it is usually much easier to compute these coefficients by a long-division expansion of the transfer function X_u/X_i, as is described in Appendix I of reference 1. If the terms of the numerator and denominator are arranged in ascending powers of s, the steady-state coefficients result from the long division; whereas if the terms are arranged in descending powers of s, the initial-value coefficients result. By equation 15, it can be shown that the lowest-order nonzero initial-value coefficient is equal to

$$a_m=s^m\frac{X_u}{X_i}\bigg|_{s=\infty}=s^mG\bigg|_{s=\infty}\quad(16)$$

and by equation 14, the lowest-order nonzero steady-state coefficient equal to

$$c_n=s^n\frac{X_u}{X_i}\bigg|_{s=0}\quad(17)$$

Assuming the loop has greater than zero gain at zero frequency, the step response coefficient c_0 is never zero, and hence is always the lowest-order nonzero steady-
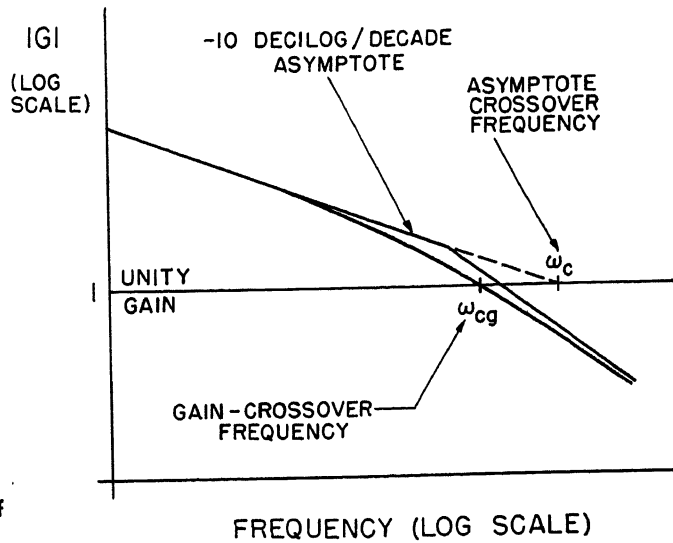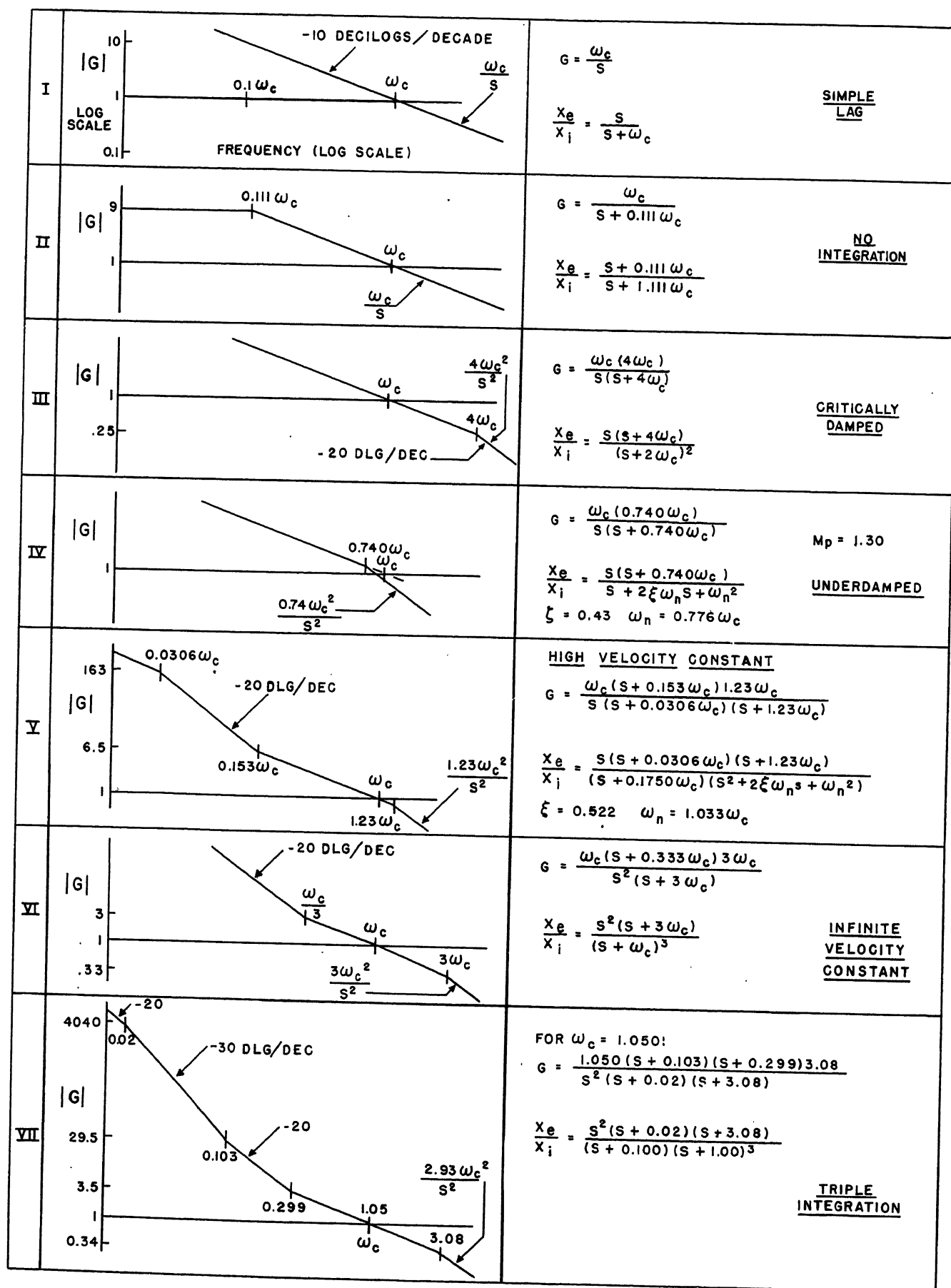
Fig. 2. Description of the systems
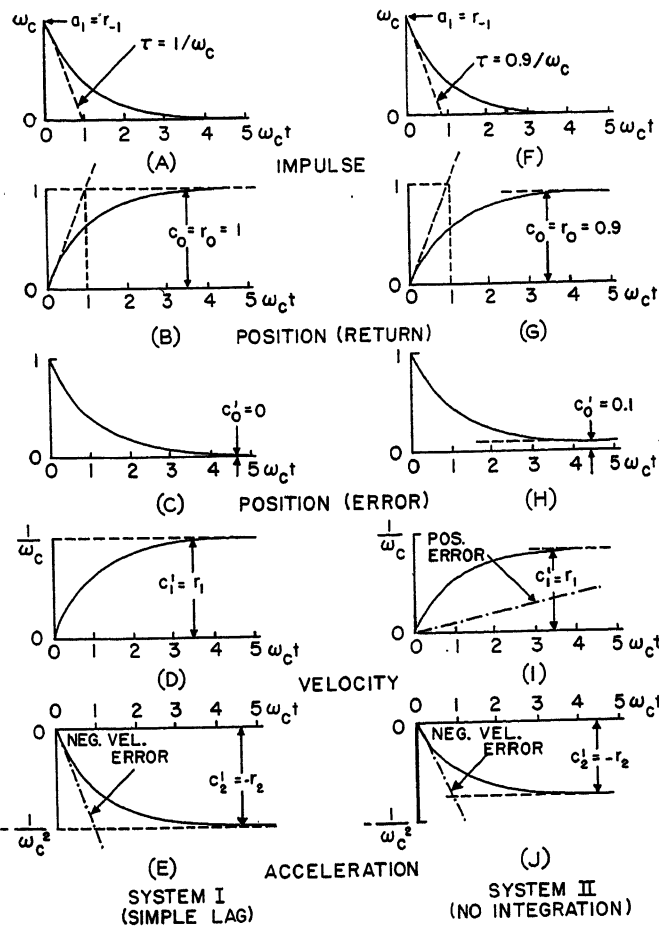
Fig. 3 (left). Transient terms for systems I and II, single-order systems

ω_c ←a_1 = r_{-1}

τ = 1/ω_c

O   0 1 2 3 4 5 ω_c t

(A)   IMPULSE

c_0 = r_0 = 1

(B)   POSITION (RETURN)

c_0' = 0

(C)   POSITION (ERROR)

$\frac{1}{\omega_c}$ c_1' = r_1

(D)   VELOCITY

O 1 2 3 4 5 ω_c t

NEG. VEL. ERROR   c_2' = -r_2

$-\frac{1}{\omega_c^2}$

(E)   ACCELERATION

SYSTEM I
(SIMPLE LAG)

ω_c ←a_1 = r_{-1}

τ = 0.9/ω_c

O   0 1 2 3 4 5 ω_c t

(F)

c_0 = r_0 = 0.9

(G)

c_0' = 0.1

(H)

$\frac{1}{\omega_c}$ POS. ERROR   c_1' = r_1

(I)

O 1 2 3 4 5 ω_c t

NEG. VEL. ERROR   c_2' = -r_2

$-\frac{1}{\omega_c^2}$

(J)

SYSTEM II
(NO INTEGRATION)

Fig. 4 (below). Transient terms for system III, critically damped system

$\omega_c^2 \times 4$ ←a_2

3
2
1
O
-1

(A) DOUBLET

ω_c

(B) IMPULSE

O 1 2 3 4 5 ω_c t

(C) POSITION

O 1 2 3 4 5 ω_c t

$\frac{1}{\omega_c}$ c_1' = r_1   (D) VELOCITY

O 1 2 3 4 5 ω_c t

c_2' = -r_2   (E) ACCELERATION

NEG. VEL. ERROR

$-\frac{1}{\omega_c^2}$

state coefficient for the return response. On the other hand, a similar relation holds as follows for the lowest-order non-zero error coefficient

$$c_n' = s^n \left. \frac{X_e}{X_i} \right|_{s=0} \qquad (18)$$

and for the error response, the step-re-

sponse coefficient c_0' is frequently zero.

The transient terms and the steady-state and initial-value coefficients do much more than describe the response to a specific series of transient inputs; they describe the response to any input. As reference 1 shows, the transient responses of a system define its general time response in a very practical manner: there is a small set of transient responses that is significant for a given system, and if all of them are known, then each one need be specified only roughly for one to determine the response to any input.

When the transient terms and the

steady-state and initial-value coefficients are employed for this set of transient responses, the response to any input can be readily constructed; and the method of construction (which is developed in reference 1) is summarized in Appendix I.

The set of transient responses that is significant for a given system is fixed, regardless of the input, and the limits on this set depend upon the characteristics of the system transfer function. As the paper examines the transient responses for the individual systems, it presents simple physical reasons why the responses to transients outside this set need not be considered; and additional reasons are presented in reference 1.

Examination of Transient Responses of the Seven Systems

Seven different types of loop transfer functions are illustrated in Fig. 2. The transient responses are given in Appendix II and are plotted in Figs. 3 through 8. Fig. 2 gives asymptotic magnitude plots of the loop gain, $|G(j\omega)|$, for each system, and gives equations for the loop transfer function G and the error transfer function X_e/X_i.

The systems are chosen so as to show the effect upon the transient responses of the following:

1. Low-frequency integrations.

2. Bandwidth (gain crossover frequency).

3. Peaking in the return frequency response.

The transient responses are nondimensional in terms of the asymptote crossover frequency ω_c. For most of the systems, ω_c is closely equal to the actual gain crossover frequency ω_{cg}; but for system IV, ω_{cg} is $0.72\omega_c$ and for system V it is $0.85\omega_c$. Hence for these two systems, some of the transient responses are also nondimensionalized in terms of ω_{cg}.

In the plots of the transients, those for the return response are plotted for steps of integrals, and those for the error response are plotted for steps of derivatives and for a step of the input. For systems I and II, the return response is also plotted for a step of the input.

Step Response

Fig. 9(A) shows the error step response with a time scale nondimensionalized in terms of ω_{cg} of all the systems except system II. The time for the error step response to fall to 37 per cent (%) of its initial value (or, in other words, the time for the return response to rise to within 63% of its final value) is called the step response rise time t_r. The figure shows

Fig. 5. Transient terms for system IV, underdamped system

Fig. 6. Transient terms for system V, high velocity constant

that the values for the rise time are roughly equal to $1/\omega_{cg}$ for all the responses. The actual bound on the value for t_r for the different transients in Fig. 9(A) is

$$t_r = \frac{1.1}{\omega_{cg}} \pm 10\% \qquad (19)$$

The approximate inverse relationship between the step response rise time and the gain crossover frequency is very general and reliable, and it is because of this that the gain crossover frequency can be considered as such a good criteria for bandwidth.

The step responses of Fig. 9(A) for systems IV, V, and VI have 25% overshoot, whereas system VII has 30%. If the magnitude of the return frequency response were plotted for each of the systems, it would show that systems IV, V, and VI have a peak magnitude, or M_p, of about 1.3, while system VII has a somewhat greater value. Thus, the plots substantiate the well-known relationship that the value of M_p essentially determines the amount of overshoot of the step response.

There has been a growing tendency in recent years to disregard this relationship between M_p and peak overshoot and, instead, to try to deduce overshoot from the damping ratio of the so-called dominant pair of closed-loop poles. It is frequently assumed that this pair of poles must be underdamped to produce overshoot. However, systems VI and VII have no underdamped poles, as is shown in Fig. 2, but yet the step responses have as much overshoot as that of system IV, for which the damping ratio is 0.43. Thus, the shapes of the frequency response plots can give more reliable information concerning peak overshoot then can the values of the closed-loop poles.

All the systems but system II have a pure integration at low frequencies and hence an infinite positional constant or, in other words, a zero positional error-coefficient c_0'. The positional error-coefficient c_0' represents the final value of the error step response, and as Fig. 9(A) shows, the error step responses for all except system II approach zero in the steady state. System II has a zero frequency gain of 9, so that its positional

constant is 10 and its positional error-coefficient c_0' is 0.10. Thus, the error response for system II to a unit step approaches a final value equal to 0.10, as is shown in Fig. 3(G).

The return-response position steady-state coefficient for system II is

$$c_0 = 1 - c_0' = 0.9 \qquad (20)$$

Thus, the return response of system II to a unit step approaches the final value 0.9, as shown in Fig. 3(F). On the other hand, for the other systems the return-response positional coefficient c_0 is unity. Hence, as is illustrated in Fig. 3(B) for system I, the return responses to a unit step for all except system II approach a final value of unity or, in other words, the return signal is equal to the input in the steady state.

Ramp Response

For System I, the velocity constant is equal to the asymptote crossover frequency ω_c, so that the velocity error cient c_1' is equal to $1/\omega_c$. Hence, the steady-state value of the error to a unit ramp should be $1/\omega_c$. Fig. 3(D) shows

the error response of system I to a unit ramp, which rises exponentially to the final value $1/\omega_c$.

Now, consider what happens to the ramp response when an integral network is added to the system to produce a high velocity constant. Consider system VI which has an infinite velocity constant, and consequently a velocity error-coefficient c_1' equal to zero. Its response to a unit ramp is shown in Fig. 7(D). Initially, the error rises almost the same as for system 1, as if it were approaching as a final value $1/\omega_c$, but then the effect of the low-frequency integration takes over and integrates the value of the error down to zero. However, the integral network does not appreciably change the maximum value of the error. For any system, the maximum value of error to a unit ramp is primarily a function of bandwidth, and is roughly equal to $1/\omega_{cg}$. Adding gain at low frequencies in a feedback-control loop can reduce the error after it has first built up, but to maintain the error small from the start requires an increase in bandwidth.

System V has an integral network, which does not produce an infinite velocity

constant, but increases the velocity constant by a factor of 5 to the value $5\omega_c$. Fig. 6(D) shows the error response of the system to a unit ramp. The maximum error is closely equal to $1/\omega_{cg}$, which is somewhat greater than $1/\omega_c$ since for system V, ω_{cg} is less than ω_c. The integral network reduces the value of error from $1/\omega_{cg}$ down to the final value $1/5\omega_c$. Comparing this ramp response with that of system VI in Fig. 7(D) shows that the ramp response of system V, besides having a nonzero final value, also has a much longer settling time. The reason for this is that for system V the upper break frequency of the integral network, which is $0.153\ \omega_c$, is less than half that for system VI, which is $\omega_c/3$. The upper break frequency of the integral network determines the speed of the integrating action, and hence the settling time of the ramp response.

System VII has an infinite velocity constant, as does system VI, but it also has a triple integration for a significant region at low frequencies, which produces a relatively high acceleration-constant, or a low acceleration-error-coefficient c_2'. Fig. 8(D) shows its error response to a unit

ramp. Although the response eventually settles to zero, the triple integration adds a long negative tail to it, thus greatly prolonging the settling time.

The ramp error responses for systems III and IV are shown in Figs. 4(D) and 5(D). The response for system III is very nearly the same as that for system I, but for system IV it is more oscillatory, because the system is underdamped. Another explanation for the shape of the ramp response for system IV, is that its maximum error is closely equal to $1/\omega_{cg}$, whereas its final value is equal to $1/\omega_c$. Since ω_{cg} is less than ω_c for system IV, the error transient must overshoot beyond the value $1/\omega_c$ before it settles.

The error response of system II to a unit ramp is shown in Fig. 3(I). A curious aspect of this response is that the velocity error coefficient c_1' is smaller than that for system I, in spite of the lower gain that system II has at low frequencies. In contrast, the responses of the other systems have indicated that the more gain at low frequencies the smaller the velocity error coefficient. The explanation to this paradox is that the ramp error response of system II is composed of two parts: an exponentially rising composite transient and a linearly increasing steady-state positional-error term. Since system II has a steady-state error proportional to the input position the response to a ramp input must have a steady-state term equal to the positional error coefficient c_0' multiplied by the input ramp. When the positional-error component is added to the exponential velocity-error component, the total error for the ramp input is at all times greater than that for system I, despite the fact that the velocity error coefficient is lower for system II.

It is also desirable to examine the return response for a ramp input as well as the error response. Fig. 10 gives the return responses of systems I and VI to a ramp segment that starts at $\omega_c t=0$ and breaks back to zero slope at $\omega_c t=8$. The return response for system I has essentially the same shape as the input, except that the corners are rounded and it lags the input by the time delay $1/\omega_c$. For system VI, there is an initial time delay at the start of the ramp nearly equal to $1/\omega_c$, but then the integrating action of the low-frequency integration takes over and reduces the steady-state time delay to zero. The integration does this by storing up a signal large enough to compensate for the steady-state velocity error and, consequently, at the end of the ramp segment the return response must overshoot the input curve and decay back slowly, as the low-frequency integration

Fig. 8. Transient terms for system VII, triple integration

discharges the stored compensating signal.

For a ramp input, the instantaneous value of time interval by which the return response lags the input is equal to the instantaneous value of the error response to a unit ramp. For example, the maximum error for system VI to a unit ramp is $0.85/\omega_c$ and, as shown in Fig. 10, the maximum time delay of the return response to a ramp is also $0.85/\omega_c$. To prove the relationship in general, consider a ramp input of velocity V

$$x_i(t) = Vt \tag{21}$$

and examine the return response, which can be expressed as

$$x_u(t) = x_i(t) - x_e(t) = Vt - x_e(t) \tag{22}$$

Evaluate $x_u(t)$ at a given time t_2 and determine the value of time t_1 at which the input, given by equation 21, is equal to this value of $x_u(t)$.

$$x_u(t_2) = Vt_2 - x_e(t_2) = x_i(t_1) = Vt_1 \tag{23}$$

This gives for the time interval $t_2 - t_1$

$$(t_2 - t_1) = \frac{x_e(t_2)}{V} \tag{24}$$

The value $(t_2 - t_1)$ represents the time interval by which the return signal at the time t_2 lags the input. The equation shows that the time delay for a ramp input (of any size) is equal to the error response to a unit ramp.

Since the maximum error to a unit ramp is roughly equal to the rise time to a step, the maximum time delay by which the ramp response lags the input is roughly equal to the step-response rise time. This approximation can be further extended to apply to an arbitrary input. The return response of system I to an arbitrary input is essentially that input delayed by the step-response rise time $1/\omega_c$, except that components that rise and fall in an interval smaller than the time delay are attenuated. For systems with additional low-frequency integrations, such as system VI, the response is not so simple, because the integrating action is constantly storing up signals to compensate for steady-state errors due to constant velocity (and sometimes acceleration) components, and adding overshoot whenever those components end. Because of this integrating action there

is not a uniform delay between the input and return signals; nevertheless, whenever the input turns a sharp corner a momentary delay is introduced which is essentially equal to the rise time of the step response.

In general, the primary aspects of the response of any low-pass transfer function is a time delay inversely proportional to its bandwidth. The delay is a fundamental property and must occur whenever abrupt changes occur in the input. On the other hand, the low-frequency portion of the transfer function can be set to take advantage of constant-velocity and possibly of constant-acceleration characteristics of the input by distorting the response wave form from that of the input, so that small errors can be achieved during periods when those derivatives are essentially constant.

It is important to recognize the limitations inherent in reducing the error of a feedback control loop by means of a low-frequency integration. When a constant-velocity component of the input commences, the integral network must wait until the error has built up before it

Biernson—Transient Responses from Open-Loop Frequency Response
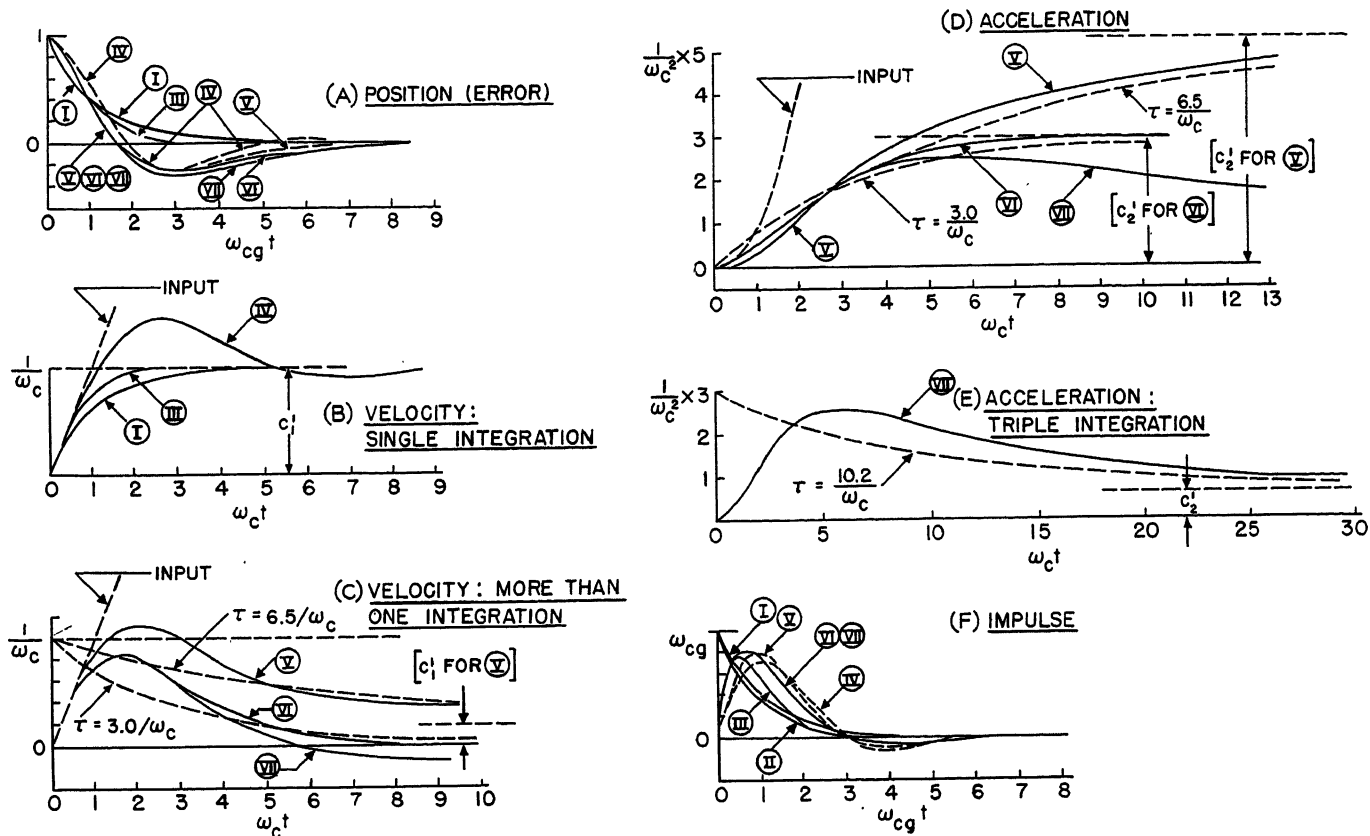
Fig. 9. Comparison of corresponding transient terms

begins to correct for it; and after that component has ended, the network must add overshoot and a tail to the response. This overshoot and tail must also occur when the system responds to a step, as is shown by the step responses of systems V, VI, and VII in Figs. 6(C), 7(C), and 8(C). Thus, a double integration can materially reduce the error of a system for an input which has significant velocity components maintained essentially constant for periods much longer than the ramp response settling-time; but if the velocity is not very constant, a double integration may even increase the resultant error. Besides, for certain applications in which the system is following a programmed input, it may be desirable that the shape of the output curve be like that of the input even though a time lag occurs between the two. Since a double integration achieves its reduction in steady-state error by distorting the output curve, a double integration is generally undesirable in such applications where a time delay can be tolerated.

RESPONSE TO STEP OF ACCELERATION

The response to a step of acceleration is an important characteristic of the dynamic response of systems V, VI and VII, which have double integrations at low frequency and hence low steady-state velocity errors. However, for the other systems, the steady-state velocity errors are so large in comparison to the acceleration components of error that the response to a step of acceleration is not very important.

For system I, the solid curve in Fig. 3(E) shows the acceleration-error portion of the response to a unit step of acceleration. Strangely enough, this error component is negative (in the opposite direction to the input), and prior to the time $\omega_c t = 1$, it is even greater in magnitude than the input. On the other hand, when the acceleration has a step, the velocity must increase as a ramp; so that there must also be a linearly increasing steady-state component of error equal to the ramp of velocity multiplied by the velocity error coefficient c_1'. This velocity-error component, although positive, is plotted in Fig. 3(E) in the negative direction for ease of comparison, and is labeled "Neg. Vel. Error," which means the negative of the velocity error. The plot shows that the velocity error component is at all times greater in magnitude than the acceleration error component, which indicates that the acceleration error is relatively unimportant.

Fig. 11(A) shows how the acceleration error component of the response of system I to a unit step of acceleration subtracts from the linearly increasing velocity component. The resultant error curve is almost the same as the velocity error, but it lags the velocity error by the time delay c_2'/c_1', which is equal to $1/\omega_c$. Fig. 11(B) gives the ramp response, which shows that the time delay of the velocity error-term produced by the acceleration error is equal to the rise time of the ramp error-response. Thus, the velocity error component is delayed by an interval equal to the rise time of the ramp error response in the same manner that the ramp response is delayed by an interval equal to the rise time of the step response.

To generalize, the effect of the acceleration component of error for an arbitrary input to a system with a monotonic ramp error response (which includes systems I, II, and III) is essentially to delay the velocity component of error by an interval equal to the rise time of the ramp error response. Thus, the acceleration component of error for these systems, in fact, for all systems with only a single integration at low frequencies, should be considered as not a basic component of error but merely a correction to be added to the velocity component of error. If the acceleration component were neglected in a computation, this would not appreciably change the maximum value of error, but only shift the time at which it occurs.
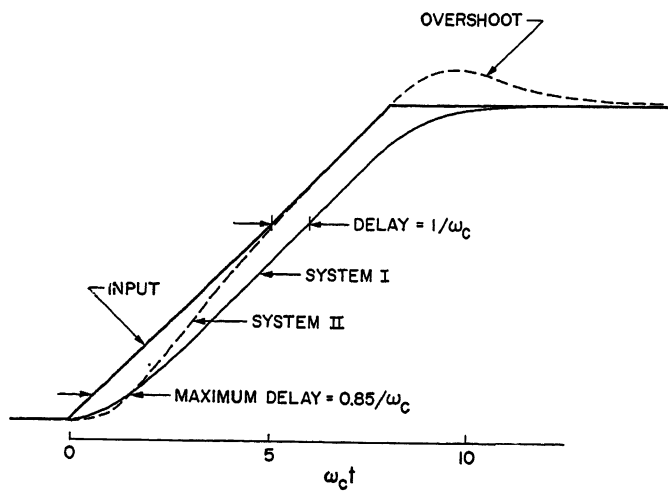
OVERSHOOT

DELAY = $1/\omega_c$

SYSTEM I

INPUT

SYSTEM II

MAXIMUM DELAY = $0.85/\omega_c$

0   5   10

$\omega_c t$

The step-of-acceleration responses of systems II and III are essentially the same as for system I, but the response for the underdamped system IV is somewhat different. However, the acceleration component of error is still quite small with respect to the steady-state velocity component (except at short values of time) and hence is of little relative importance. The underdamped nature of the poles of system IV makes its response more difficult to estimate than for the other systems. Every time system IV is disturbed, it tends to oscillate in a rather independent manner. However, if the system is reasonably well damped, these oscillations do not greatly affect the response, but they do limit the accuracy achievable by rough approximations.

The acceleration component of error is of little importance for the systems with one integration. However, when a double integration is employed for a significant frequency region below gain crossover, to achieve a high velocity constant, the steady-state velocity errors are so reduced that the acceleration errors become quite important in comparison. Consequently, for systems V, VI, and VII, the response to a step of acceleration is very significant.

The error response of system VI to a unit step of acceleration is shown in Fig. 7(E). It rises monotonically to a final value equal to the acceleration error coefficient and the curve along which it rises has essentially the same shape as the tail of the ramp response in Fig. 7(D). Since the velocity error-coefficient for system VI is zero, there is no velocity error term in its response to a step of acceleration.

The response of system V to a step of acceleration, shown in Fig. 6(E), has a linearly increasing velocity error term along with the acceleration error term. However, the velocity error term is small with respect to the acceleration error, ex-

cept after a rather long period; which indicates that for an arbitrary input to system V acceleration components of error can be appreciable.

The response of system VII to a step of acceleration, shown in Figs. 8(E) and 8(F), demonstrates the manner in which the triple integration produces a low acceleration error coefficient. The error builds up almost to the value of the acceleration error coefficient that would exist without the third integration, but then the effect of the third integration slowly reduces the error to a final value equal to the actual acceleration error coefficient c_2'. The settling time of the transient is a function of the upper break frequency of the third integration (at 0.103 in Fig. 2), which must be set much lower than the gain crossover frequency in order to achieve adequate stability. Consequently, the settling time of the response of a triple-integration system to a step of acceleration must be considerably longer than the rise time of the step response. In fact the settling time is so long for the step-of-acceleration response of system VII that it had to be replotted in Fig. 8(F) with a condensed time scale, to show the manner in which it settles.

Generally, a triple integration is of little use in reducing steady-state errors in a feedback control application. Although it can decrease the acceleration error coefficient without requiring that the bandwidth be increased, this results in a reduction of error only for acceleration components maintained for periods of time which are very long in comparison to the step-response rise time. Although velocity components of the input to a feedback control system are often maintained constant for periods considerably longer than the step-response rise time, a large constant-acceleration component of long duration is quite rare. When a triple integration occurs in a feedback

control loop it generally is not included for the purpose of decreasing the acceleration error coefficient. Sometimes, for example, a triple-integration characteristic is encountered because an integral network is added to a system to increase its stiffness in opposing load torques.

An important fault of the triple integration is that it adds a long-negative tail to the ramp error response, as was shown in Fig. 8(D), which greatly prolongs the ramp response settling time. Because of this effect, a triple integration often results in an increase in low-frequency errors.

RESPONSE TO STEP OF RATE OF ACCELERATION

The responses of the single-integration systems to steps of rate of acceleration are of no importance, and hence are not presented. The responses of the double-integration systems, systems V and VI, to steps of rate of acceleration are shown in Fig. 6(F) and 7(F). The figures show that the rate-of-acceleration error components are small in comparison with the linearly increasing steady-state acceleration error terms, except at short values of time and, hence, the rate-of-acceleration error should be considered as merely a correction upon the acceleration error. Figure 12(A) shows how the rate-of-acceleration component of error for the response of system VI, to a step of rate of acceleration, subtracts from the acceleration error term. The resultant error is essentially the acceleration error delayed by the time interval c_3'/c_2', which is equal to $2.6/\omega_c$. This time delay is approximately equal to the rise time of the error response to a step of acceleration, which is shown in Fig. 12(B) to be $2.9/\omega_c$. Thus, for a double-integration system, the essential effect of the rate-of-acceleration-component of error is to delay the acceleration component of error; and the time interval of this delay is roughly equal to the rise time of the acceleration error portion of the response to a step of acceleration.

The response of system V to a step of rate of acceleration, given in Fig. 6(F), shows only two of the three error components. Besides the composite rate-of-acceleration transient and the linearly increasing acceleration term which are shown, the response also contains a parabolic velocity error term, equal to the velocity error coefficient multiplied by the resultant parabola of velocity.

For system VII, the triple integration reduces the acceleration error coefficient sufficiently to make the response to a step of rate of acceleration quite significant. This response, which is shown in
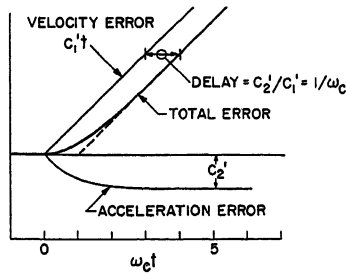
(a) ERROR RESPONSE TO STEP OF ACCELERATION
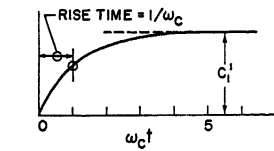
(b) ERROR RESPONSE TO RAMP

Fig. 11. Effect of acceleration component of error of system I

(A) ERROR RESPONSE TO STEP OF RATE-OF-ACCELERATION

(B) ERROR RESPONSE TO STEP OF ACCELERATION

Fig. 12. Effect of rate-of-acceleration component of error of system VI

Fig. 8(G), consists of two portions, a composite rate-of-acceleration transient and a linearly increasing acceleration error term. The composite transient rises very slowly, with essentially the same settling time as the tail in the step-of-acceleration response of Fig. 8(F).

STEP OF FOURTH DERIVATIVE

System VII is the only one for which the response to a step of the fourth derivative is significant, and even then it is not very important. The error response is plotted in Fig. 8(H) and consists of three portions: a composite transient portion, a linear rate-of-acceleration error portion, and a parabolic acceleration error portion. The two steady-state error components are positive, but are plotted in the negative direction. For a general input, the essential effect of the fourth-derivative component of error is to delay the rate-of-acceleration component of error by a time interval equal to c_4'/c_3', which is $14/\omega_c$ and is roughly equal to the rise time of the rate-of-acceleration transient.

IMPULSE RESPONSE

To determine the response of a feedback control system to high-frequency inputs, particularly to noise inputs, the impulse response should be examined, and possibly the responses to a doublet, a triplet, etc. These represent the responses to steps of integrals of the input: an impulse is a step of the first integral; a doublet is a step of the second integral, etc. As has been shown, the maximum number of integrations below gain crossover determines how many derivative step responses are significant. Similarly, the maximum number of integra-

tions above gain crossover, that is, the maximum steepness of the magnitude slope at high frequencies, determines how many integral step responses are significant.

By equation 16, the initial value of the impulse response of the systems, which represents the initial-value coefficient a_1, is equal to

$$a_1 = sG|_{s=\infty} \qquad (25)$$

For systems I and II, the loop transfer function G approaches the asymptote ω_c/s at high frequencies, and hence the initial-value coefficient a_1 is equal to ω_c. The transfer functions of the other systems approach double integrations at high frequency, and hence for them the initial-value coefficient a_1 is zero. Thus, the impulse response for systems I and II starts at the value ω_c, as shown in Fig. 3(A) and 3(F); whereas, for the other systems the impulse response starts at zero as is shown in section (B), Figs. 4 through 8.

In Fig. 9(F) the impulse responses of all the systems are nondimensionalized in terms of the gain crossover frequency ω_{cg}. This comparison shows that, regardless of the initial value of the responses, the maximum values of all of them are essentially equal to ω_{cg}. The peak value of the impulse response is a measure of the amount a system is disturbed by noise, because an impulse acts as a noise input to a system. Thus, systems with the same ω_{cg} may be considered to have the same bandwidth with respect to their filter action upon noise signals because an impulse disturbs all of them by roughly the same amount.

The concept of bandwidth has two aspects: a system follows frequency components below the bandwidth, and filters those at frequencies above. The rise time of the step response is a good indication of the low-frequency aspect, because it describes how fast the system follows the low-frequency components of the input; and the maximum value of the impulse response is a good indication of the high-frequency aspect because it is a measure of how much the system is disturbed by unwanted high-frequency noise. The frequency ω_{cg} thus passes both tests as a criterion for bandwidth. Systems with the same ω_{cg} have approximately the same rise time, and have approximately the same peak value for the impulse response.

It is possible for the maximum value of the impulse response of a system to be much greater than ω_{cg}. This occurs when the loop gain has very poor attenuation above ω_{cg}, which could be caused by high-frequency open-loop zeros. As might be
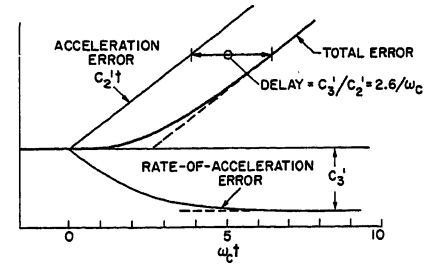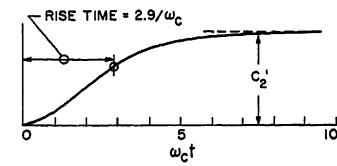
expected, such a system is much more bothered by noise than a system with the same gain crossover frequency and more rapid high-frequency attenuation, but it does not follow the input signal any better. It has a larger impulse response, but the rise time of the step response is no faster.

DOUBLET RESPONSE

For systems I and II, the doublet response includes an impulse of magnitude a_1. Although this response has some significance in determining the system response to general high-frequency inputs, it is relatively unimportant in comparison with the portion associated with the impulse response initial-value coefficient a_1. Consequently, the doublet response was not plotted for these two systems.

Since the coefficient a_1 is zero for systems III to VII, the doublet response initial-value coefficient for these systems can be calculated from

$$a_2 = s^2 G|_{s=\infty} \qquad (26)$$

This value can be computed by inspection, by determining from Fig. 2 the equation for the high-frequency asymptote. For example, the high-frequency asymptote of system V is $1.23\,\omega_c^2/s^2$, so that the initial-value coefficient a_2 must be equal to $1.23\,\omega_c^2$. This is equal to the initial value of the doublet response as shown in Fig. 6(A). The doublet responses of systems III to VII are shown in part (A) of Figs. 4 through 8.

TRIPLET RESPONSE

The triplet responses of systems I and II are of no interest but those of the other systems are somewhat significant. For systems I and II, the triplet response

System	c_4	c_1'	c_2'	c_3'	c_4'	a_1	a_2	a_3
I	0	$1/\omega c$	$-1/\omega c^2$			ωc	$-\omega c^2$	
II	0.10	$0.81/\omega c$	$-0.729/\omega c^2$			ωc	$-1.111\omega c^2$	
III	0	$1/\omega c$	$-3/4\omega c^2$			0	$4\omega c^2$	$-16\omega c^3$
IV	0	$1/\omega c$	$0.346/\omega c^2$			0	$0.740\omega c^2$	$-0.547\omega c^3$
V	0	$1/5\omega c$	$5.35/\omega c^2$	$-32.0/\omega c^3$		0	$1.23\omega c^2$	$-1.36\omega c^3$
VI	0	0	$3/\omega c^2$	$-8/\omega c^3$		0	$3\omega c^2$	$-8\omega c^3$
VII	0	0	$0.690/\omega c^2$	$26.6/\omega c^3$	$-374/\omega c^4$	0	$2.93\omega c^2$	$-7.53\omega c^3$

contains an impulse of magnitude a_2 and a doublet of magnitude a_1; while the triplet response for the other systems contains just an impulse of magnitude a_2. For an arbitrary high-frequency input to the systems, the portion of the response associated with the triplet transient is usually small in comparison with the portion because of the initial-value coefficient a_2, but it may still be significant.

Rules for Approximating Transients

By comparing the transient responses of the systems, general rules for approximating transient response can be induced. An important basis of many of these approximation rules is the relationship between the open-loop frequency response and the closed-loop poles.

VALUES OF CLOSED-LOOP POLES

Since the values of the closed-loop poles determine the time constants of the various transient components, an approximate knowledge of the closed-loop poles is necessary for estimating the shapes of the transient responses. As shown in reference 2, the closed-loop poles of a feedback control loop are roughly equal to the following:

1. The open-loop zeros under ω_c in frequency.

2. The open-loop poles over ω_c in frequency.

3. Plus a pole roughly equal to $-\omega_c$.

The approximation rules to be presented will apply only to the step response and to the response to steps of derivatives; since these responses are generally much more important in feedback control calculations than the responses to steps of

Table II. Transient Coefficients for System V

n	K_{1n}'	$|2K_{2n}'|$	φ_{2n}'
-2	$-0.00507\omega c^2$	$2.88\omega c^2$	$-57.3°$
-1	$0.0290\omega c$	$2.79\omega c$	$-178.8°$
0	-0.166	2.70	$59.7°$
1	$0.949/\omega c$	$2.61/\omega c$	$-61.8°$
2	$-5.42/\omega c^2$	$2.53/\omega c^2$	$176.7°$
3	$30.9/\omega c^3$	$2.46/\omega c^3$	$55.2°$

integrals. The shapes of the transients of these responses are determined primarily by the mid-frequency and low-frequency poles. However, it is not necessary to know the values of the mid-frequency poles when estimating the shapes of the transients, because the important effects of these poles can be determined from the gain crossover frequency and peaking (M_p) of the frequency response. Thus, it is the low-frequency poles that must be estimated, and as shown in the foregoing, item 1, these are roughly equal to the low-frequency zeros of the open-loop transfer function.

The actual low-frequency closed-loop poles are shifted somewhat from the open-loop zeros, and the per-unit magnitude of shift from a particular zero is roughly equal to the reciprocal of the loop gain, i.e., the magnitude of G, at the break frequency of that zero. In other words, the greater the loop gain, the smaller the shift. On the other hand, fairly good approximations can be obtained by ignoring the shifts and assuming that the low-frequency closed-loop poles are actually equal to the low-frequency open-loop zeros. This assumption will serve as a basis for developing rules for approximating transient response.

DEVELOPMENT OF APPROXIMATION RULES

In the following paragraphs, the transients of the systems are re-examined in order to determine ways of approximating them.

Ramp Response

The unit-ramp error responses of the systems are compared in Figs. 9(B) and 9(C). For all the systems, the error follows the input fairly closely until the error reaches a maximum value approximately equal to $1/\omega_{cg}$ and, after that, the error decays to a value equal to the velocity error coefficient. For systems with one pure integration, the velocity error coefficient is equal to $1/\omega_c$; and since ω_c is generally fairly closely equal to ω_{cg}, the final value of error is roughly equal to the maximum value.

Fig. 9(C) shows the ramp error re-

sponses for the systems with more than one integration. For these systems, the final value of error is considerably less than the maximum value. Since system V has a low-frequency zero with the break frequency $0.153\ \omega_c$, it must have a closed loop pole with roughly the same break frequency; and this pole must produce a component in the transient response with a time constant roughly equal to $1/0.153\ \omega_c$, or $6.5/\omega_c$. In Fig. 9(C) a dashed curve, representing an exponential with this time constant, is drawn from the value $1/\omega_c$ to a final value equal to the velocity error coefficient c_1' for system V. The figure shows that this exponential is a reasonably good approximation of the tail portion of the ramp response of system V. Thus, the value of the low-frequency open-loop zero defines in a rather quantitative manner the shape of the tail of the ramp response.

For systems VI and VII, the first open-loop zero below gain crossover has a break frequency roughly equal to $\omega_c/3$, and hence the tail of the ramp error response of each system should roughly follow an exponential with the time constant $3/\omega_c$. Such an exponential is shown by a dashed curve in Fig. 9(C) decaying from $1/\omega_c$ to zero. The response of system VI follows this exponential fairly well as it decays, but that of system VII has an additional negative exponential, which causes the response to overshoot and to have a long negative tail. This additional exponential is due to the closed-loop pole shifted from the low-frequency zero with the break frequency 0.13 (or about $0.1\omega_c$). Thus, the triple integration adds to the ramp response of system VII a small but long negative exponential, with a time constant roughly equal to $1/0.1\omega_c$ or $10/\omega_c$.

Table III. Transient Coefficients for System VI

n	K_{1n}	K_{2n}	K_{3n}
-2	$3\omega c^2$	$-5\omega c^3$	ωc^4
-1	0	$3\omega c^2$	$-\omega c^3$
0	1	ωc	$-\omega c^2$
1	0	1	ωc
2	$-3/\omega c^2$	$-3/\omega c$	-1
3	$8/\omega c^3$	$5/\omega c^2$	$1/\omega c$

Table IV. Transient Coefficients for System VII

n	K_{1n}	K_{2n}	K_{3n}	K_{4n}
-2	$2.94\omega_c^2$	$-4.74\omega_c^3$	$0.929\omega_c^4$	$0.000311\omega_c^2$
-1	$0.00311\omega_c$	$2.93\omega_c^2$	$0.973\omega_c^3$	$-0.00311\omega_c$
0	0.967	$0.922\omega_c$	$-1.023\omega_c^2$	0.0327
1	$0.343/\omega_c$	1.293	$1.075\omega_c$	$-0.343/\omega_c$
2	$-4.34/\omega_c^2$	$-3.73/\omega_c$	-1.13	$3.66/\omega_c^2$
3	$11.20/\omega_c^3$	$6.51/\omega_c$	$1.18/\omega_c$	$-37.8/\omega_c^3$
4	$-21.8/\omega_c^4$	$-9.35/\omega_c^3$	$-1.27/\omega_c^2$	$400/\omega_c^4$

Step-of-Acceleration Response

Fig. 9(D) shows the composite tran-' sient portions of the responses of systems V, VI, and VII to a unit step of accelera- tion. The composite transient for system V is very closely an exponential with the time constant $6.5/\omega_c$, rising to a final value equal to its acceleration error coeffi- cient c_2'. Similarly, that for system VI is nearly an exponential with the time con- stant $3/\omega_c$ rising to a final value equal to its acceleration error coefficient. For sys- tem VII, the composite transient rises in essentially the same manner as does that for system VI, and almost reaches the value $3/\omega_c^2$, which is the acceleration error coefficient for system VI and is the error coefficient that system VII would have if the triple integration were not present. However, the triple integration finally causes the transient to decay along an exponential with a time constant roughly equal to $10/\omega_c$, to a final value equal to the actual acceleration error coefficient.

Step of Rate-of-Acceleration Response

The composite transient portion of the response of system VII to a step of rate of acceleration, shown in Fig. 8(G), ap- proximates an exponential with the time constant $10/\omega_c$ rising to a final value equal to the rate-of-acceleration error coefficient.

Step Response

The step error responses of all but sys- tem II are shown in Fig. 9(A). All the step error responses, including that of system II, fall to within 37% of the final value in a time roughly equal to $1/\omega_{cg}$. The amount of overshoot of the step re- sponse is primarily a function of the M_p, which represents the peak magnitude of the return (or closed-loop) frequency re- sponse. For an M_p of 1.3, there is roughly 25% overshoot in the step re- sponse. For the high-velocity-constant systems, V, VI and VII, the second inte- gration adds to the step response a nega- tive exponential component with a time constant roughly equal to the upper break frequency of that integration. This ex- ponential accounts for a large part of the

overshoot of the step response, and adds a tail. For system VII, the third integra- tion also adds a very small positive tail to the step error response. This tail is of too small an amplitude to be seen on the plots; nevertheless, it may be quite im- portant because of its long duration. Since the upper break frequency of the third integration is approximately $\omega_c/10$, this tail has a time constant of roughly $10/\omega_c$, which is 10 times greater than the step-response rise time.

RELATION BETWEEN SUCCESSIVE TRANSIENTS

As was shown in equation 13, the tran- sient terms of successive derivatives are related by integration. Since the coeffi- cient c_n is equal to $-T_n(0)$, this relation can be expressed as a definite integral as the following

$$T_n(t) + c_n = \int_0^t T_{n-1}(t)dt, \text{ for } n \geq 0 \qquad (27)$$

Thus, the integral of a transient term is the composite transient term for the next higher derivative. This relation can be used to explain important characteristics of certain transient responses.

The equation shows that the ramp error response is the integral of the step error response. Consequently, for the final value of the ramp error response to be much smaller than its maximum value, the net integral of the step error re- sponse must be roughly zero or, in other words, the negative area must be roughly equal to the positive area. Thus, to achieve a high velocity-constant, relative to the bandwidth, the step response must overshoot. For a given step-response rise time, the positive area of the step error response is essentially fixed (unless the system is so oscillatory that there is a great deal of undershoot) and this fixes the overshoot area that must exist if the system is to have a high velocity-constant. Since the area of the overshoot is fixed, the step response of a high-velocity-con- stant system must either have a large peak overshoot or a long tail. The length of the tail can be adjusted by changing the upper break frequency of the second

integration. However, if the step-re- sponse tail is lengthened to reduce the peak overshoot, the tail on the ramp re- sponse is lengthened by the same amount, and the system must wait longer to achieve the effect of the high velocity- constant.

Similarly, if the system employs a triple integration to achieve a decrease in the acceleration error coefficient, the ramp response must overshoot and the area of overshoot must be roughly equal to the positive area (as is shown by the ramp response of system VII in Fig. 8(D). Thus, with the triple integration, there must either be a large overshoot or a long tail in the ramp error response; and a long tail in the ramp response requires an equally long tail in the step-of-accelera- tion response, which causes a long delay before the benefit from the decrease in acceleration error coefficient may be achieved.

Conclusions

The paper has considered a great many of the important frequency-response characteristics of feedback control sys- tems and has shown how they affect the shapes of the transient responses and the response to more general inputs. Besides a general discussion of the main effects, it has presented a number of rules for ap- proximating the transient response by means of inspecting the frequency-re- sponse plots.

SUMMARY OF APPROXIMATION RULES

The rules for approximating the tran- sients of feedback control systems for a unit step and for units steps of deriva- tives are summarized as follows.

Step Response

The rise time of the step response, defined as the time to reach within 37% of the final value, is roughly equal to $1/\omega_{cg}$.

The peak overshoot is determined by the value of M_p, there being about 25% over- shoot for an M_p of 1.30.

If the loop has a pure integration the steady-state error is zero; if not, it is equal to the positional error coefficient.

A second low-frequency integration produces in the step error response a nega- tive exponential tail with a time constant equal to the reciprocal of the upper break frequency of that integration. A third integration adds a smaller but longer positive exponential tail with a time constant equal to the reciprocal of the upper break frequency of the third inte- gration.

Ramp Response

For a system with one pure integration, the ramp error response approximately follows the linearly rising input until reaching a maximum error of $1/\omega_{cg}$, and after that it approaches a final value equal to $1/\omega_c$. If the system does not have a pure integration, the response is essentially the same, except that there is an additional linear error term equal to the positional error coefficient multiplied by the ramp input curve.

For a system with two integrations, the ramp error response rises the same as for one integration to a maximum error of $1/\omega_{cg}$, but then decays exponentially to the actual value of the velocity error coefficient. The time constant for the exponential decay is equal to the reciprocal of the upper break frequency of the second integration.

For a system with three integrations the ramp error response is the same as for a system with two integrations, except that there is an additional negative exponential producing a long tail, which has a time constant equal to the reciprocal of the upper break frequency of the third integration.

Response to Step of Acceleration

For a system with two integrations, the composite acceleration transient is a rising exponential, with a time constant equal to the reciprocal of the upper break frequency of the second integration, and a final value equal to the acceleration error coefficient.

For a system with three integrations, the composite acceleration transient initially rises as if the system had only two integrations, and nearly reaches an error equal to the value of the acceleration error coefficient the system would have without the third integration. However, the third integration finally causes the error to decay exponentially from this maximum value to the actual value of acceleration error coefficient; and the time constant of the exponential is equal to the reciprocal of the upper break frequency of the third integration.

Response to Step of Rate of Acceleration

For a system with three integrations, the composite rate-of-acceleration transient is a rising exponential with a time constant equal to the reciprocal of the upper break frequency of the third integration and a final value equal to the rate-of-acceleration error coefficient.

When these rules refer to systems with two or three integrations, it does not imply that the integrations must hold at zero frequency. A system with two inte-

grations is one that has an open-loop magnitude slope of -20 decilogs per decade for a significant frequency region below gain crossover; while a system with three integrations is one with a -30 decilog-per-decade slope for a significant frequency region. A system with two integrations in general has one open-loop zero below ω_c in frequency, and the break frequency of this zero represents the upper break frequency of the second integration. Similarly, a system with three integrations has two open-loop zeros below ω_c, and the break frequency of the higher is the upper break frequency of the second integration, while that of the lower is the upper break frequency of the third integration.

GENERALITY OF APPROXIMATIONS

The rules for approximation have been presented as having general applicability but, of course, their general applicability has not been actually proved. Nevertheless, sufficient justification has been presented for their use in engineering design. After all, the well-known relationship between M_p and the peak overshoot of the step response has never been proved, but it has long been a very useful design tool.

On the other hand, the author is now preparing a paper describing a general successive-approximation method for computing the transient responses, in which the gross computation is based upon the values of the frequency response asymptotes. The method provides a frame of reference for developing approximations relating frequency response to transient response; and since the method is capable of extension to achieve high accuracy, the designer has a ready means for evaluating the adequacy of the approximations. The method will help to justify the approximations presented in this paper but, more important, it will provide more adequate rules of approximation and a broader explanation of the dynamic principles of feedback control.

In a particular application the designer can evaluate the accuracy of an approximation by calculation or by an analogue computer study. Nevertheless, such techniques cannot substitute for the approximations, because they are far too complex to provide an adequate frame of reference for intelligent design. If the design problem is reasonably complex and there are a great many alternatives to choose from, there is no substitute for a basic comprehension of the important dynamic aspects involved and such a comprehension must be based mainly upon reliable approximations.

Appendix I. Summary of Method for Computing Response to Arbitrary Input

This Appendix summarizes the technique described in reference 1 for calculating the response of a system to an arbitrary input. The technique is based upon the expansion of the system response to a step of a derivative and to a step of an integral of the input.

The expansion of the return response to a step of the nth derivative of the input which was given in equation 2, can also be expressed as

$$x_u(t) = c_0 x_i(t) + c_1 D x_i(t) + c_2 D^2 x_i(t) + \ldots + [c_n + T_n(t)] D^n x_i(t) \quad (28)$$

while the response to a step of the mth integral, which was given in equation 3, can be expressed as

$$x_u(t) = a_0 x_i(t) + a_1 D^{-1} x_i(t) + \ldots + a_{m-1} D^{-(m-1)} x_i(t) + T_{-m}(t) D^{-m} x_i(t) \quad (29)$$

The symbol $D^n x_i(t)$ represents the nth time derivative of the input $x_i(t)$, and the symbol $D^{-m} x_i(t)$ represents the mth time integral.

The expansions show that the response to an arbitrary input can be calculated as described in the following. First, break the input down as follows into a sum of steps of derivatives and integrals of the input.

1. Approximate as steps those portions of the input which rise more abruptly than the system step response.
2. Draw an average curve through high-frequency portions which rise and fall significantly faster than the system step response.
3. With the steps and high-frequency portion removed, differentiate the remaining smooth curve.
4. Approximate with steps those portions of the derivative curve which rise more abruptly than the transient term for that derivative, and differentiate the remaining smooth curve.
5. Execute step 4 repeatedly until reaching a specified derivative, and at that point completely approximate the curve by straight-line segments. Since the derivative of this approximation is entirely composed of steps, making this approximation terminates the procedure. The order of the derivative at which the final straight-line approximations are made is a function only of the system transfer function and does not depend upon the input (provided the input is not so simple that it is completely approximated at a lower derivative).
6. The high-frequency portion of the input, which was separated, should be repeatedly integrated and, for each integral, step approximations should be made to subtract out any constant bias components which are longer in duration than the rise time of the corresponding system transient term.
7. At a specified integral, which also depends only upon the system transfer function, the integration process should be stopped, and the effects of higher integrals neglected.

Then form the response curve, by summing together the following three portions:

1. The derivative curves multiplied by the corresponding steady-stage coefficients.
2. The integral curves multiplied by the corresponding initial-value coefficients.
3. For each step of derivative or integral a corresponding transient term proportional to the size of the step.

The accuracy of the calculated response can be defined quite adequately by a simple bound. The inaccuracy caused by approximating a particular derivative or integral curve can be no greater than the product obtained by multiplying the maximum deviation between the exact and approximate curves by the corresponding remainder coefficient. The remainder coefficient r_n for the nth derivative is equal to

$$r_n = \int_0^\infty |T_{n-1}(t)| \, dt \qquad (30)$$

and can be calculated easily by a graphical construction upon the transient term $T_n(t)$. For the mth integral, the remainder coefficient is designated r_{-m} and is calculated in a similar manner from $T_{-m}(t)$.

An important aspect of this method is that there are very definite stopping points in the differentiation and integrative processes, which depend only upon the system transfer function, and therefore for a given system there is only a fixed set of transient responses which need be known. The limits on this set are as follows:

The highest derivative transient response needed is the first one for which the composite transient term $[c_n + T_n(t)]$ is no greater in magnitude than the term $c_{n-1}t$ due to the lower-order steady-state coefficient. The highest integral transient response needed is the one beyond the first transient for which the initial value is its maximum value in magnitude.

The differentiation process is terminated by completely approximating the highest derivative curve by steps, which is performed by making straight-line approximations on the lower-order derivative curve. At this lower-order derivative, the error coefficient is generally nearly equal in magnitude to the remainder coefficient, so that the per-unit accuracy of the calculated error response is essentially as good as the per-unit accuracy of the straight-line approximation.

Appendix II. Calculation of Elements of Transient Responses of the Seven Systems

This Appendix lists all the steady-state and initial-value coefficients and the transient terms for those responses of the seven systems that are plotted in the figures.

Coefficients

The values of the steady-state and initial-value coefficients of the seven systems are listed in Table I. These were computed by long division expansions of the return transfer functions, following the procedure described in Appendix I of reference 1.

Transient Terms for Systems V Through VII

The equations for the transient terms for systems V, VI, and VII are listed in the following. Those of the other systems are derived in the next section to illustrate the procedure for computing transient terms.

System V

The form of the error-response transient terms is

$$T_n'(t) = K_{1n}\epsilon^{-0.175\omega_c t} + |2K_{2n}| \epsilon^{-\zeta\omega_n t} \times \sin(\omega_0 t + \phi_n) \quad (31)$$

where

$$\zeta\omega_n = 0.540\omega_c \qquad (32)$$

$$\omega_0 = 0.880\omega_c \qquad (33)$$

The coefficients of this transient are listed in Table II. This system is the same as that used for illustration in reference 1, and the computation of the transient terms is presented in Appendix II of that reference.

System VI

The form of the error-response transient terms is

$$T_n'(t) = \epsilon^{-\omega_c t}(K_{1n} + K_{2n}t + K_{3n}t^2) \qquad (34)$$

The coefficients are listed in Table III.

System VII

The form of the error-response transient terms is

$$T_r'(t) = \epsilon^{-0.95\omega_c t}(K_{1n} + K_{2n}t + K_{3n}t^2) + K_{4n}\epsilon^{-0.095\omega_c t} \qquad (35)$$

The coefficients are listed in Table IV.

Transient Terms for Systems I Through IV

The transient terms for systems I through IV are calculated as follows.

System 1

For system I, the error transfer function is, from Fig. 2

$$\frac{X_e}{X_i} = \frac{s}{s + \omega_c} \qquad (36)$$

For a unit step-of-position input, $X_i(s)$ is $1/s$, and $X_e(s)$ is equal to the position transient term $T_0'(s)$.

$$X_e(s) = T_0'(s) = \frac{1}{s}\frac{X_s(s)}{X_i(s)} = \frac{1}{1 + \omega_c} \qquad (37)$$

The inverse transform is

$$x_e(t) = T_0'(t) = \epsilon^{-\omega_c t} \qquad (38)$$

Equation 11 has shown that for single-order poles, the nth derivative transient term can be expressed as

$$T_n'(t) = \sum_{l=1}^{L} K_{1n}\epsilon^{s_l t} = \sum_{l=1}^{L} \frac{K_{l0}}{(s_l)^n}\epsilon^{s_l t} \qquad (39)$$

where K_{l0} is the step-response coefficient for the pole s_l and K_{ln} is the corresponding coefficient for the nth derivative. Hence,

the general equation for the nth-derivative transient term for system I is

$$T_n'(t) = \frac{\epsilon^{-\omega_c t}}{(-\omega_c)^n} = (-1)^n\frac{\epsilon^{-\omega_c t}}{\omega_c^n} \qquad (40)$$

System 2

$$\frac{X_e}{X_i} = \frac{s + 0.111\omega_c}{s + 1.111\omega_c} \qquad (41)$$

For a unit step of position input

$$X_e(s) = \frac{c_0'}{s} + T_0'(s) = \frac{(s + 0.11\omega_c)}{s(s + 1.11\omega_c)}$$
$$= \frac{0.10}{s} + \frac{0.90}{s + 1.11\omega_c} \qquad (42)$$

The inverse transform is

$$x_e(t) = c_0' + T_0'(t) = 0.1 + 0.9\epsilon^{-1.11\omega_c t} \qquad (43)$$

The general equation for the nth-derivative transient term is, then

$$T_n'(t) = \frac{0.9\ \epsilon^{-1.11\omega_c t}}{(-1.11\omega_c)^n} = (-1)^n 0.9 \times \left[\frac{0.9}{\omega_c}\right]^n \epsilon^{-1.11\omega_c t} \qquad (44)$$

System 3

$$\frac{X_e}{X_i} = \frac{s(s + 4\omega_c)}{(s + 2\omega_c)^2} \qquad (45)$$

For a step-of-position input

$$X_e(s) = T_0'(s) = \frac{(s + 4\omega_c)}{(s + 2\omega_c)^2}$$
$$= \frac{1}{(s + 2\omega_c)} + \frac{2\omega_c}{(s + 2\omega_c)^2} \qquad (46)$$

The inverse transform is

$$T_0'(t) = (1 + 2\omega_c t)\epsilon^{-2\omega_c t} \qquad (47)$$

Since this system has a double-order pole, equation 39 does not apply. Instead one must use the more general relation of equation 13, which applies to all transfer functions, which is

$$T_n(t) = \int T_{n-1}(t)dt \qquad (48)$$

Thus, the derivative transient terms are obtained by integration and the integral transient terms obtained by differentiation. To integrate, use the following integration equations.

$$\int \epsilon^{-u}du = -\epsilon^{-u} \qquad (49)$$

$$\int u\epsilon^{-u}du = -\epsilon^{-u}(1 + u) \qquad (50)$$

$$\int u^2\epsilon^{-u}du = -\epsilon^{-u}(2 + 2u + u^2) \qquad (51)$$

Equations 49 and 50 yield the derivative transient terms

$$T_1'(t) = \int T_0'(t)dt = \frac{-\epsilon^{-2\omega_c t}}{\omega_c}(1 + \omega_c t) \qquad (52)$$

$$T_2'(t) = \int T_1'(t)dt = \frac{3}{4\omega_c^2}\epsilon^{-2\omega_c t}\left(1 + \frac{2\omega_c t}{3}\right) \qquad (53)$$

Equation 51 is required for a system with a triple-order pole. Differentiate to obtain the integral transient terms.

$$T_{-1}(t) = -T_{-1}'(t) = \frac{-d}{dt} T_0'(t)$$

$$= (2\omega_c)(2\omega_c t)\epsilon^{-2\omega_c t} \quad (54)$$

$$T_{-2}(t) = \frac{d}{dt} T_{-1}(t) = (2\omega_c)^2 \epsilon^{-2\omega_c t}(1-2\omega_c t) \quad (55)$$

SYSTEM 4

$$\frac{X_e}{X_i} = \frac{s(s+0.740\omega_c)}{s^2+2\zeta\omega_n s+\omega_n^2} \quad (56)$$

where $\zeta = 0.43$ and $\omega_n = 0.860\ \omega_c$. For a unit step of position $X_e(s)$ is

$$X_e(s) = T_0'(s) = \frac{K_0}{(s+\zeta\omega_n-j\omega_0)} + \frac{K_0^*}{(s+\zeta\omega_n+j\omega_0)} \quad (57)$$

where

$$K_0 = \frac{\omega_n}{2\omega_0}\underline{/\theta-90 \text{ degrees}} \quad (58)$$

$$\theta = \cos^{-1}\zeta \quad (59)$$

$$\omega_0 = \omega_n \sqrt{1-\zeta^2} \quad (60)$$

The value of the upper-half-plane pole is

$$-\zeta\omega_n+j\omega_0 = -\omega_n\underline{/-\theta} \quad (61)$$

Hence, the nth derivative coefficient K_n is equal to, by equation 39

$$K_n = \frac{K_0}{(-\zeta\omega_n+j\omega_0)^n} = \frac{K_0}{(-\omega_n\underline{/-\theta})^n} \quad (62)$$

Substitute into equation 62 the expression for K_0 in equation 58

$$K_n = \frac{(-1)^n(\omega_n)^{1-n}}{2\omega_0}\underline{/(n+1)\theta-90 \text{ degrees}} \quad (63)$$

The nth-derivative transient term is

$$T_n'(s) = \frac{K_n}{s+\zeta\omega_n-j\omega_0} + \frac{K_n^*}{s+\zeta\omega_n+j\omega_0} \quad (64)$$

Take the inverse transform, substituting the expression for K_n in equation 63

$$T_n'(t) = \frac{(-1)^n(\omega_n)^{1-n}}{\omega_0} \times \epsilon^{-\zeta\omega_n t} \sin\left[\omega_0 t+(n+1)\theta\right] \quad (65)$$

Substituting the values for the parameters gives

$$T_n'(t) = \frac{(-1)^n 1.108}{(0.860\omega_c)^n} \epsilon^{-0.37\omega_c t} \times \sin\left[0.776\omega_c t+(n+1)64.5 \text{ degrees}\right] \quad (66)$$

References

1. A SIMPLE METHOD FOR CALCULATING THE TIME RESPONSE OF A SYSTEM TO AN ARBITRARY INPUT, G. A. Biernson. *AIEE Transactions*, vol. 74, pt. II, September 1955, pp. 227–45.

2. QUICK METHODS FOR EVALUATING THE CLOSED-LOOP POLES OF FEEDBACK CONTROL SYSTEMS, George Biernson. *AIEE Transactions*, vol. 72, pt. 2, May 1953, pp. 53–70.

3. A SIMPLE METHOD FOR CALCULATING THE TIME RESPONSE OF A SYSTEM TO AN ARBITRARY INPUT, G. A. Biernson. *Report no. 7138-R-3*, Servomechanisms Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., Jan. 20, 1954. ASTIA AD-25082.

Discussion

Rufus Oldenburger (Woodward Governor Company, Rockford, Ill.): Dr. LeRoy A. MacColl of the Bell Telephone Laboratories has said that there are a number of simple relations between frequency and transient response in use at the Bell Laboratories. In particular, Dr. MacColl found that the build-up time of a transient is about $1/2f_0$ for the cutoff frequency f_0. For a normally attenuating magnitude ratio curve the cutoff frequency is the frequency at which the magnitude ratio is down to half of the value it has for low frequencies. The frequency of the oscillations for an oscillatory transient response is about equal to the cutoff frequency f_0. The rate of decay of the ripple amplitude depends on how fast the magnitude ratio frequency response curve drops. There appears to be some overlapping between the results of the Bell Laboratories and those of the author. However, as far as I know, the Bell work has not been published.

Simple approximate relations between frequency and transient responses are invaluable for automatic control work. Qualitative design on paper is necessary before one goes into fine details such as those which might be applied to a computor. The author is to be congratulated for his excellent contribution to this subject.

George A. Biernson: In Mr. Oldenburger's constructive discussion he refers to the important MacColl approximation. There are a number of approximations in general use in the field of feedback control, but many of them are so inadequate that there is a general skepticism of the accuracy achievable by reliable approximations. That by MacColl represented a good early step in the process of relating frequency response and transient response. Now, however, much more reliable approximations exist.

MacColl's approximation relates the magnitude of the return (or closed-loop) frequency response to the build-up time of the step response. In contrast, the author has related the magnitude of the loop (or open-loop) frequency response to the rise time of the step response. The author's approximation is not only simpler to apply than the approximation of MacColl, but much more reliable and exact. This conclusion may seem strange to many, because the literature often leads to the belief that the closed-loop (return) transfer function must be considered in order to determine the performance of the over-all closed loop. MacColl's approximation is reasonably accurate only for certain types of transfer functions, whereas the author's is extremely general and appears to hold for all reasonable feedback control systems.

One of the difficulties of the MacColl approximation is that it describes the "build-up time" which evidently is the time for the unit step response to rise to unity, and as such applies only to step responses with a significant amount of overshoot. The parameter used by the author is the "rise time," which is the time for the step response to rise to 63% of its final value. Since this parameter applies to overdamped as well as to underdamped responses, it is of much more general applicability. The "settling time," which is the time for the response to settle to within a small given per cent of its final value, might also be considered, but because the low-frequency characteristics of the loop transfer function can produce tails in the step response, the "settling time" cannot generally be considered as merely a function of bandwidth.

A much more severe difficulty of the Mac-Coll approximation is that it is based upon the magnitude, rather than the phase, of the return frequency response. The build-up time of the step response represents a time delay, and as such corresponds to a phase lag in the frequency domain. If in a feedback control loop there is a

reasonable amount of peaking M_p of the magnitude of the return frequency response and the loop gain increases at a reasonably rapid rate at frequencies below gain cross-over, the following three frequency parameters are roughly equal to each other.

1. The gain crossover frequency ω_{cg}.
2. The frequency of 60 degrees of phase lag of the return frequency response.
3. The frequency where the magnitude of the error frequency response rises to 0.707, i.e., $1/\sqrt{2}$.

Any of these parameters may be considered as a good measure of the bandwidth for feedback control applications. The step-response rise time is roughly equal to the reciprocal of the value of any one of them in radians per second.

The purpose of a feedback control system is to follow the reference input (or a calibrated function of that input) with small error. Just because the magnitude curve of the return frequency response is reasonably flat up to a given frequency does not mean that the system can follow frequency components up to that frequency with reasonably small error. At frequencies where the phase lag of the return frequency response has exceeded 60 degrees, the smaller the magnitude of the return response, the smaller the error. Thus, for ideal response, the magnitude should attenuate rapidly at frequencies above the point of a 60-degree phase lag. Any response where the phase lag is greater than 60 degrees merely contributes to the error.

If the return transfer function is a simple quadratic, the approximation of MacColl is equivalent to that of the author. For a quadratic, the cut-off frequency f_0, where the magnitude is down to one-half, is related to the frequency f_{60}, where the phase lag is 60 degrees, by

$$f_0 = 1.9f_{60} \pm 10\%$$

The accuracy limit applies for all values of

damping ratio less than unity. Thus, MacColl's approximation for the build-up time is equivalent to the following for a quadratic

$$t_{bu} = \frac{1}{2f_0} = \frac{1}{2(1.9f_{60})} = \frac{2\pi}{3.8\omega_{60}} = \frac{1.65}{\omega_{60}}$$

where ω_{60} is the frequency f_{60} in radians per second. Since the rise time t_r is the time for the step response to rise to 63%, it is roughly equal to 63% of t_{bu}, if the step response has a reasonable overshoot. This gives for the rise time

$$t_r = 0.63t_{bu} = \frac{1.05}{\omega_{60}}$$

Thus, for a quadratic the approximation of

Dr. MacColl agrees with that of the author.

On the other hand, there are a great many practical feedback control systems in which the return frequency response does not approximate a simple quadratic. In particular, feedback control loops which contain sampling devices, such as magnetic amplifiers, generally produce a return frequency response with magnitude greater than unity significantly beyond the point of 180 degrees of phase lag. For such loops, the approximation of MacColl is extremely poor.

Textbooks and other literature on feedback control have generally overemphasized the importance of the plot of the magnitude of the return frequency response. The measure of bandwidth is usually determined from this plot, and it is generally assumed that this measure defines the speed of response to a step. Thus, the criticism the author is making of MacColl's approximation represents a general criticism of much of the thinking in the field on this matter of bandwidth.

An important practical difficulty of defining bandwidth in terms of the magnitude plot of the return frequency response is that it often leads the designer to try to improve the response of a system by extending the magnitude plot, without regard to the phase curve. The author once observed an engineer painfully at work to increase the resonant frequency of a feedback control loop, only to find that after he had increased it by more than a factor of 2, the step response was even slower than when he started.

Analysis of Single-Phase-to-3-Phase Static Phase Converters

J. C. HOGAN
MEMBER AIEE

A STATIC phase converter is a device which supplies polyphase power to a load from a single-phase power source using only passive circuit elements. Ideally such a device should produce balanced polyphase voltages at the load. Several recent articles[1,2,3] have been published concerning the operation of 3-phase induction motors connected through static phase converters to single-phase lines. These articles discuss the simple form of phase converter using only capacitors. This capacitor-only type of phase converter can be adjusted to give balanced 3-phase power only for loads possessing a lagging power factor of 0.5. For any other load it is impossible to obtain balanced conditions. In contrast, the Add-A-Phase phase converter, consisting of an autotransformer in addition to the capacitors, may be set to supply balanced 3-phase power to any load possessing a lagging power factor regardless of its value. This more general type of phase converter circuit will be referred to in this article as the autotransformer type of phase converter. At the point at which balanced operation occurs with an induction motor load, the speed, torque, horsepower, efficiency, and other motor quantities will be identical to those obtained when the motor is connected to a balanced 3-phase power source of the same voltage.

It is the purpose to present an analysis of the autotransformer phase converter,

treating the capacitor-only phase converter as a special case of the more general analysis. The analysis indicates that many of the limitations placed on the application of the capacitor-only phase converter may be relaxed when the autotransformer phase converter is used.

Analysis

The general form of the circuit for phase converters is given in Fig. 1. The voltage at terminal A' may be varied according to the tap setting on the autotransformer. The capacitor X_c in series with phase A may also be varied to obtain balanced conditions. This circuit may be analyzed by the method of symmetrical components by considering the voltages at A', B', and C' as a set of unbalanced 3-phase voltages applied to a circuit with unequal line impedances going to a 3-phase motor. This set of voltages may be expressed in terms of the single-phase voltage applied to the phase converter and the tap setting of the autotransformer. Thus

$$V_{A'B} = NV_1 \tag{1}$$

$$V_{BC} = V_1 \tag{2}$$

$$V_{CA'} = -(1+N)V_1 \tag{3}$$

where

$$N = \frac{|V_{A'B}|}{|V_{BC}|} \tag{4}$$

The positive negative and zero-sequence component voltages are given by

$$V_{A'B_1} = \frac{1}{3}(V_{ab} + aV_{bc} + a^2V_{ca}) \tag{5}$$

$$V_{A'B_2} = \frac{1}{3}(V_{ab} + a^2V_{bc} + aV_{ca}) \tag{6}$$

$$V_{A'B_0} = \frac{1}{3}(V_{ab} + V_{bc} + V_{ca}) \tag{7}$$

Substituting equations 1 to 3 into equations 5 to 7 yields

$$V_{A'B_1} = \frac{V_1}{\sqrt{3}}(1\underline{/90^\circ} + N\underline{/30^\circ}) \tag{8}$$

$$V_{A'B_2} = \frac{V}{\sqrt{3}}(1\underline{/-90^\circ} + N\underline{/-30^\circ}) \tag{9}$$

$$V_{A'B_0} = 0 \tag{10}$$

Since the sum of the three currents must also equal zero, the zero-sequence component of the current is also zero.

The equivalent circuit of the induction motor is usually expressed in terms of an equivalent wye circuit, with the applied voltage being a line-to-neutral voltage. Expressing the component voltages of equations 8 and 9 in terms of line-to-neutral voltages gives

$$V_{A'n_1} = \frac{V_{A'B_1}}{\sqrt{3}}\underline{/-30^\circ} \tag{11}$$

$$V_{A'n_1} = \frac{V_1}{3}[N\underline{/0^\circ} + 1\underline{/60^\circ}] \tag{12}$$

$$V_{A'n_2} = \frac{V_{A'B_2}}{\sqrt{3}}\underline{/30^\circ} \tag{13}$$

$$V_{A'n_2} = \frac{V_1}{3}[N\underline{/0^\circ} + 1\underline{/-60^\circ}] \tag{14}$$
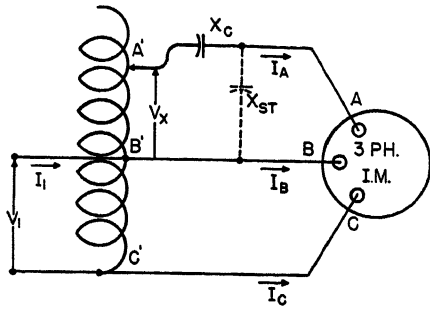
Fig. 1. Circuit of autotransformer type of phase converter

Next the unsymmetrical line imped-ances must be resolved into a form suit-able for application of the sequence rule. The sequence components of these unbal-anced line impedances are

$$Z_{A0} = \frac{1}{3}(Z_A + Z_B + Z_C) \quad (15)$$

$$Z_{A1} = \frac{1}{3}(Z_A + aZ_B + a^2Z_C) \quad (16)$$

$$Z_{A2} = \frac{1}{3}(Z_A + a^2Z_B + aZ_C) \quad (17)$$

In the phase-converter circuit

$$Z_B = Z_c = 0 \quad (18)$$

Thus equations 15, 16, and 17 become

$$Z_{A0} = Z_{A1} = Z_{A2} = \frac{Z_A}{3} \quad (19)$$

The equations of the circuit may now be written in terms of the sequence com-ponents. Thus

$$V_{an1} = I_{a1}(Z_{A0} + Z_{m1}) + I_{a2}Z_{A2} \quad (20)$$

$$V_{an2} = I_{a1}Z_{A1} + I_{a2}(Z_{A0} + Z_{m2}) \quad (21)$$

where

I_{a1}, I_{a2} = the positive- and negative-sequence component currents going to the motor respectively.
Z_{m1}, Z_{m2} = the positive- and negative-sequence impedances per phase of the induction machine respectively.

Substituting equations 12, 14, and 19 into equations 20 and 21 yields
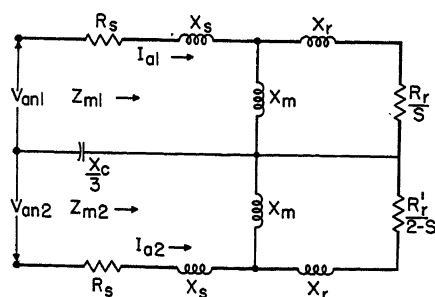
Fig. 2. Equivalent circuit of phase converter and induction motor load

$$\frac{V_1}{3}(N\underline{/0^\circ} + 1\underline{/60^\circ}) = I_{a1}\left(\frac{Z_A}{3} + Z_{m1}\right) + I_{a2}\frac{Z_A}{3} \quad (22)$$

$$\frac{V_1}{3}(N\underline{/0^\circ} + 1\underline{/-60^\circ}) = I_{a1}\frac{Z_A}{3} + I_{a2}\left(\frac{Z_A}{3} + Z_{m2}\right) \quad (23)$$

Solving for the sequence currents gives

$$I_{a1} = \frac{V_1}{3}\left(\frac{Z_{m2}(N\underline{/0^\circ} + 1\underline{/60^\circ}) + \frac{Z_A}{\sqrt{3}}\underline{/90^\circ}}{Z_{m1}Z_{m2} + \frac{Z_A}{3}(Z_{m1} + Z_{m2})}\right)$$

$$(24)$$

and

$$I_{2a} = \frac{V_1}{3}\left(\frac{Z_{m1}(N\underline{/0^\circ} + 1\underline{/-60^\circ}) - \frac{Z_A}{\sqrt{3}}\underline{/90^\circ}}{Z_{m1}Z_{m2} + \frac{Z_A}{3}(Z_{m1} + Z_{m2})}\right)$$

$$(25)$$

The ideal operating condition for a 3-phase motor is to have balanced 3-phase voltages applied to the terminals of the motor which should result in balanced line currents. Thus, the negative-se-quence current must equal zero. From equation 25 it is seen that this condition is satisfied by the relationship

$$Z_{m1}/\theta_{m1}(N\underline{/0^\circ} + 1\underline{/-60^\circ}) = \frac{Z_A\underline{/90^\circ}}{\sqrt{3}} \quad (26)$$

where θ_{m1} is the positive-sequence power-factor angle of the machine. Z_A may be any type of impedance but to minimize losses it is usually a capacitor for induc-tive loads. Thus, for an induction motor load

$$Z_A = -jX_c \quad (27)$$

Since both the autotransformer tap set-ting N and the capacitive reactance X_c may be varied, it is possible to satisfy equation (26) for any polyphase induction motor and thus insure balanced opera-tion at that particular load. As the load requirements change, the motor will no longer have exactly balanced currents. It is, therefore, important to adjust initially the values of N and X_c so as to give balanced operation for the particular load condition most usually encountered. Solution of equation 26 gives

$$X_c = \frac{3}{2}\frac{|Z_{m1}|}{\operatorname{Sin}\theta_{m1}} \quad (28)$$

and

$$N = \frac{\operatorname{Cos}(\theta_{m1} + 30^\circ)}{\operatorname{Sin}\theta_{m1}} \quad (29)$$
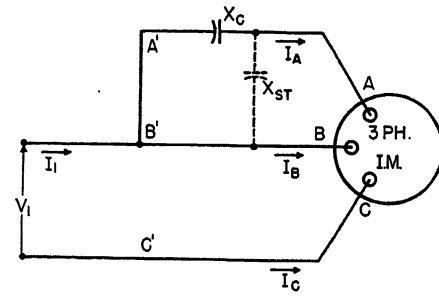
Fig. 3. Circuit of capacitor-only type of phase converter

Equivalent Circuit

Equations 22 and 23 are the equations of the equivelent circuit shown in Fig. 2. It is interesting to note that this circuit is identical in form to the equivalent circuit for a capacitor motor.[4] The circuit for the capacitor motor represents 2-phase symmetrical component currents and voltages while the circuit shown in Fig. 2 gives 3-phase quantities. This similarity in equivalent circuits indicates that the unbalance caused by negative-sequence currents in a phase converter is the same as the unbalance that occurs in capacitor-run motors caused by changing load conditions.

Capacitor-Only Phase Converter

The theory of the capacitor-only phase converters may be treated as a special case of the autotransformer phase con-verter. Adjusting the tap setting A' to point B' makes voltage $V_{A'B'}$ equal to zero or for this case

$$N = 0 \quad (30)$$

The autotransformer has no effect on the operation of the phase converter and the circuit is the same as that used in the capacitor-only phase converter as given in Fig. 3. The negative-sequence cur-rent must again be zero to obtain balanced conditions, as specified by equation 26. For the capacitor-only phase converter, this becomes
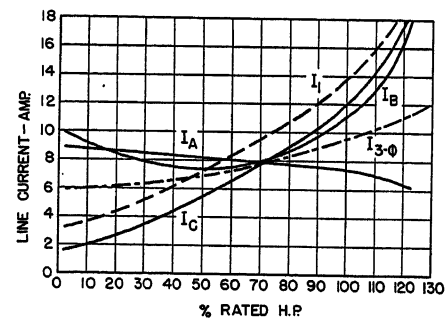
Fig. 4. Current variation caused by load changes for phase converter balanced at 70-per-cent load
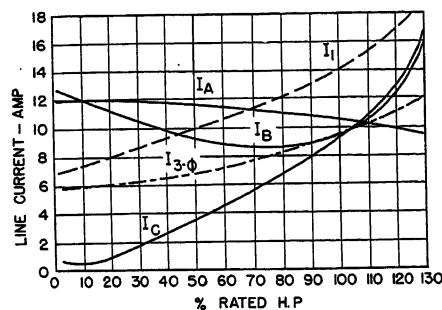
Fig. 5. Current variation caused by load changes for phase converter balanced at 110-per-cent load

$$Z_m \underline{/\theta_{m1} - 60°} = \frac{X_c}{\sqrt{3}} \underline{/0°} \tag{31}$$

This equation can only be satisfied if

$$\theta_m = 60° \tag{32}$$

and

$$X_c = \sqrt{3}|Z_{m1}| \tag{33}$$

From equation 33 it is seen that only the special case of power factor equal to 0.5 can be balanced. This, however, usually occurs only when the machine is very lightly loaded. Thus it may be concluded that the autotransformer is very necessary in phase converters to obtain the desired balanced operation for a motor operating under load conditions.

The same type of analysis discloses that a capacitor-run motor with identical windings in quadrature can only give balanced currents if the motor power factor is 0.707. By making the number of turns on the auxiliary winding different than the number of turns on the main winding, it is possible to obtain conditions for loads with power factors other than 0.707. This change in number of turns for the capacitor-run motor is equivalent to the change in voltage $V_{A'B}$ possible with the autotransformer phase converter and accomplishes the same purpose.

Operating Characteristics

To illustrate the operating characteristics of the autotransformer type of phase converter, experimental test results are given. A 3-phase 220-volt 60-cycle-per-second induction motor was used with a rating of 3 horsepower at 1,165 rpm. Load runs were made with the motor connected directly to a 3-phase line and with the motor connected to a single-phase line through the phase converter. Separate runs were made for the phase converter adjusted to give balanced conditions. Results will be given for the two runs taken when the phase converter was set to give a balance at 70 and 110 per cent of rated output. Since it is desired

to have balanced line currents to the motor, an examination of the variation of line current as a function of horsepower output is shown in Figs. 4 and 5. It is seen that the line currents can be balanced at particular values of load, although there is considerable variation in current unbalance as the load increases or decreases away from the point at which balance occurs. Curves given by Habermann[1] show that the currents may never become balanced with a capacitor-only phase converter.

The torque variation as a function of speed is of course the most important factor in matching a motor to a given load. The motor operated through the phase converter has a speed-torque characteristic very similar to the motor operated directly on 3-phase up to loads of about 120 per cent of rated horsepower, as shown in Fig. 6. For loads above this value the speed drops very rapidly and the breakdown torque is much lower than for 3-phase case. For the run connection the phase converter was set to give balanced conditions at 110 per cent of rated load. For certain types of load occasionally requiring excessive torques, it may be necessary to use the next larger size motor.

Voltage variation caused by line drop is a serious problem on rural single-phase lines. Since the voltage drop is proportional to the line current, it is desirable to keep the line current for a given power at a minimum. This is the same as saying that the power factor should be as close to unity as possible. Fig. 7 points out the excellent single-phase power-factor characteristics of the autotransformer phase converter compared to the power factor of the 3-phase motor.

The over-all efficiency on single-phase using the phase converter is almost the same as the motor efficiency on balanced 3-phase; the only losses present in the phase converter being the core and copper losses of the autotransformer. However, additional losses occur in the motor for unbalanced operation because of negative-sequence losses. This will result in over-all efficiencies approximately equal to the efficiency on 3-phase at the points of balance. As the unbalance increases, the over-all efficiency will decrease because of the additional motor losses, as shown in Fig. 8.

To provide the required starting torque, it is necessary to have an additional capacitor X_{ST} in the circuit for starting. This capacitor tends to balance the currents at starting and, after the motor comes up to speed, the capacitor X_{ST} is disconnected by means of a relay circuit
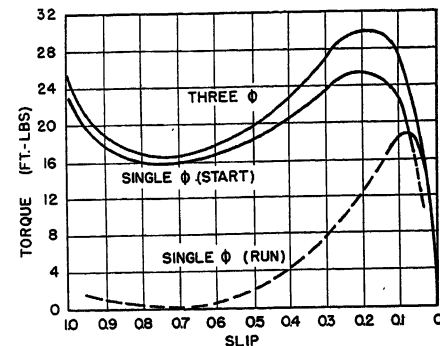
Fig. 6. Speed-torque curves for balanced 3-phase and phase-converter operation

sensitive to voltage V_{AB}. The characteristics during starting and build-up are given in Fig. 6. A tabulation of locked-rotor quantities is given in Table I. Characteristics of 3-horsepower 6-pole single-phase capacitor-start–capacitor-run motors are also given for comparative purposes. It is seen that the autotransformer phase-converter motor combination has the lowest (most desirable) value of kilovolt-amperes per foot-pound at starting. It should be pointed out that the autotransformer is just as essential to the circuit at starting as it is during running conditions. Recent tests[2] on capacitor-only phase converters indicated starting kilovolt-amperes per foot-pound of 0.86, compared to 0.57 for the auto-transformer phase converter adjusted to balance at 70 per cent of rated load. If Table I had compared 4-pole motors instead of 6-pole motors, the per-cent torques throughout the whole table would be greater than those given.

Conclusions

From this analysis of the autotransformer phase converter, the following advantages may be cited:

1. It is the only static phase converter available whose circuit is capable of providing balanced 3-phase currents to a 3-phase induction motor from a single-phase line.

2. The starting kilovolt-amperes required are greatly reduced below that of the 3-
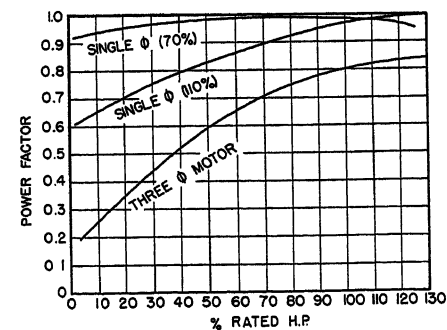
Fig. 7. Power-factor variation caused by load changes
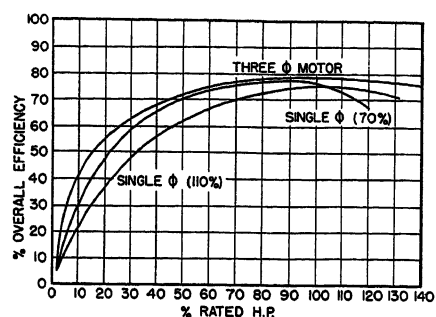
Fig. 8. Efficiency variation caused by load changes

Table I. Comparison of Operating Characteristics of 3-Horsepower Induction Motors

| | 3-Phase Motor on 3 Phases | 3-Phase Motor with Add-A-Phase | | Single-Phase Motor† |
		70-Per-Cent Balance	110-Per-Cent Balance	
Line volts	220	220	220	230
Breakdown torque, per cent of rated torque*	250	138	145	200
Starting current, amperes	48	53	58	70
Starting torque, per cent of rated torque*	210	155	195	175
Starting kilovolt-amperes	18.3	11.7	12.8	16.1
Starting kilovolt-amperes per foot-pound	0.645	0.57	0.49	0.68

*Rated torque = 13.5 foot pounds
†Capacitor-start–capacitor-run

phase motor on balanced 3-phase power or that of a single-phase motor.

3. Power factor is improved throughout the complete operating range of the motor.

4. The starting torque is comparable to that of a single-phase motor but only about 85 per cent as much as the 3-phase motor.

5. The starting kilovolt-amperes per foot-pound are smaller than those of the 3-phase or single-phase motor.

6. Cost of autotransformer phase converter plus 3-phase induction motor is competitive with the cost of a single-phase motor.

7. Simplicity of 3-phase motor reduces maintenance required.

8. Where 3-phase power may eventually become available, it permits immediate use of 3-phase motors.

9. Phase-converter units of a given horsepower rating may be used with motors smaller than that rating.

The advantages of the phase converter are as follows:

1. Breakdown torque is less than that of the 3-phase induction motor on balanced 3-phase power or that of a single-phase motor.

2. The starting torque is about 85 per cent of the starting torque of the 3-phase motor and is about the same as the starting torque for a single-phase induction motor.

It should be mentioned that 3-phase motors will increase in usage in rural areas because of the 3-phase power source that will soon be supplied on farm tractors. The 3-phase motor with a phase converter can thus be used on the single-phase lines serving the farm and, in case of emergency, the 3-phase motor can be connected directly to the portable 3-phase power source.

In conclusion, the autotransformer phase converter—3-phase induction motor combination—can be used as a source of power where only single phase is available without the usual excessive voltage drop during starting. For loads where the power required is almost constant, the unit can be adjusted to give balanced operation at a high power fac-

tor for that particular load without any derating of the motor. Where danger of explosion exists, the phase converter may be placed in a position remote from the 3-phase motor. Power companies might also permit motors to be used with higher ratings than the usual 5 to 7.5 horsepower when such motors are used in conjunction with a phase converter because of the low kilovolt-amperes taken during starting.

References

1. SINGLE-PHASE OPERATION OF A 3-PHASE MOTOR WITH A SIMPLE STATIC PHASE CONVERTER, R. Habermann, Jr. *AIEE Transactions*, vol. 73, pt. IIIB, Aug. 1954, pp. 833–37.

2. AN EVALUATION OF THE PHASE CONVERTER-THREE PHASE MOTOR, K. A. Harkness, J. W. Hooper. *Bulletin*, Kansas State College, Manhattan, Kans., circular no. 8, 1954.

3. PERFORMANCE OF THE HENRY PHASE CONVERTER, W. H. Knight, P. P. Baker, H. E. Hattrup. *Agricultural Engineering*, St. Joseph, Mich., Feb. 1954.

4. EQUIVALENT CIRCUITS OF ELECTRIC MACHINERY (book), G. Kron. John Wiley and Sons, Inc., New York, N. Y., 1951, pp. 83–84.

Discussion

J. C. McKee (Mississippi State College, State College, Miss.): The desirability of securing satisfactory performance from a polyphase motor when operated on a single-phase supply has long been recognized, and a number of methods have been proposed in the literature. Dr. Hogan's excellent symmetrical component analysis of the autotransformer-capacitor method gives a firm basis on which decisions, concerning the application of such devices, may be based.

A similar symmetrical component analysis and evaluation was done concurrently at Mississippi State College, and the results completely bear out the findings of this paper. In addition, the autotransformer impedance was considered; and its small leakage reactance was determined to have but slight effect on the capacitance required to balance the machine at a given load. Load tests were performed on a 5-horsepower motor operated by a laboratory setup with variable capacitance and a continuously variable autotransformer, and on a 10-horsepower motor operated from the commercial

autotransformer-connected unit. In both tests, at any given slip (and thus motor power factor) excellent voltage balance was obtained with adjustment of both transformer ratio and capacitance. As analysis shows, the test proved that a small change in the power-factor results in a considerable change in the current unbalance factor, I_{a1}/I_{a2}. For this reason, it was deemed advisable that balance adjustments be made at full-load slip; so that overheating effects, resulting from the change in unbalance factor as load is decreased, are somewhat offset by the decrease in load. In effect, at slips less than full-load slip, a derating is applied to the machine. Oscillograms of these machines starting under load display a decided "cushioned-start" effect which, as is discussed in this paper, is gained at the expense of starting torque when compared with its starting performance as a line-start motor. This starting characteristic is most desirable on a lengthy single-phase line, and it is likely that application will be most frequent in such situations.

Since the performance of any of the so-called "phase-converters" is completely dependent upon the motor load character-

istics, it should be considered a motor starting and operating device. As such it may perform its function very satisfactorily. Until such time that balance is made automatic, under changing load conditions, a utility serving customer equipment with 3-phase service from such a device could easily find itself with an intolerable voltage unbalance.

J. C. Hogan: The discussion by Dr. McKee is a valuable addition to this paper. In particular, the balancing of the phase converter is an important consideration in the proper application of this device. As pointed out in the paper, when the load on the motor is constant the phase converter should be set to give balanced operation at that particular load. However, as Dr. McKee points out, the balance adjustment should be made at full load when the motor is subject to widely varying loads. The effects of balancing at full load versus balancing at some other load (such as 70 per cent) for the particular motor tested is given in Figs. 4 and 5. It is gratifying to know that the results and conclusions of Dr. McKee confirm the work of the author.

A Transistorized Overvoltage Relay

N. F. SCHUH
ASSOCIATE MEMBER AIEE

PERHAPS the most important single protective device in any aircraft electric system is the overvoltage relay. Overvoltage conditions cannot only cause serious damage to connected load equipment and to the power generation equipment, if not removed, but can cause the loss of the entire electric power system. Overvoltage relays are designed to have inverse time-voltage characteristics because most of the damage caused by overvoltage is because of overheating. To obtain overvoltage protection and still not get nuisance operations on normal system transients, the relay must operate very rapidly on extreme overvoltages and relatively slowly on slight overvoltages. Upon completion of the time delay the relay contacts close the trip coil of the generator control relay, thus de-energizing the faulty system.

General

OPERATING LIMITS

Fig. 1 shows the operating limits for the overvoltage relay on a modern a-c electric power system as defined in reference 1. It can be seen that the relay must not operate at all below 125 volts, line-to-neutral, and must operate below 133 volts. At 140 volts the relay must operate in 0.09 to 0.6 second, at 150 volts in 0.055 to 0.26 second, and in less than 0.1 second at 180 volts. The dash lines were added by the author.

PRESENT PRACTICE

There are several means of obtaining an inverse time-voltage characteristic such as is required. The most commonly used overvoltage relay in aircraft power systems today makes use of the dashpot principle. Here the voltage being sensed is applied through a coil to a magnetic circuit containing a plunger which is closely fitted into a closed tube. The ampere-turns because of the applied voltage cause the plunger to move so as to close the air gap of the magnetic circuit. When the plunger travel is complete, the contacts close and open the generator excitation circuit. The higher the voltage, the faster the plunger moves, and, therefore the shorter the time delay. If the high voltage is a normal transient, the plunger stops before it reaches the end of its travel and then returns to its starting point. The most commonly used device for obtaining an inverse time-voltage characteristic in utility work makes use of an induction disk-type relay. Here the disk armature rotates at a speed proportional to the applied voltage and the desired characteristic is obtained.

Both of these types of relays have distinct disadvantages when applied to aircraft power systems. The major undesirable feature is their dependence on moving mechanical parts for a time delay. This dependence means that the relay designer must be certain that his design is not adversely affected by vibration, shock, acceleration, or even mounting position. In addition, he must be concerned about stray magnetic fields influencing his magnetic circuit and he must seal his plunger-type relay against altitude effects. In his attempt to meet the operational requirements under all the extreme environmental conditions encountered in aircraft applications, the designer often ends up with either a large, heavy device or a complicated device built to very close tolerances. Neither of these have the desired reliability.

STATIC DEVICES

The obvious solution to the problem of designing an overvoltage relay unaffected by these extreme environmental conditions is to design a static device. A static device making use of electronic tubes is possible, but this approach has three disadvantages which render it undesirable. The use of electronic tubes in aircraft primary power systems is unpopular from a reliability standpoint; a warm-up time is required before operation and, because of the power requirements, this type of design would be relatively large and heavy. A new overvoltage relay design is described herein which has a static time-delay circuit and which meets all the requirements of aircraft operation while remaining small and reliable.

New Relay Circuit

NEW COMPONENTS

The availability of silicon diodes and transistors has now made possible the design of a small, reliable overvoltage relay which has none of the disadvantages associated with mechanical or electronic relays. A group of these small components has been assembled into a circuit having the characteristics desired of an overvoltage relay. The purpose of this paper is to describe this new circuit, discuss some of the design problems encountered, and show what kind of operating characteristics can be expected from the relay. Some of the silicon devices have certain unique characteristics which are utilized in the design and should be described separately before the full circuit is described.

Fig. 2 shows the characteristic of a typical low-power silicon diode. Of particular interest in this design is the reverse voltage characteristic. This type of diode exhibits a very sharp Zener breakdown, i.e., there is a saturation current of only a few microamperes until the reverse voltage reaches the breakdown value. At this point the reverse current suddenly increases and is limited only by the resistance of the circuit. This breakdown characteristic makes the Zener diode a very useful device for sensing voltage levels. The same effect may be obtained by using a battery with a voltage equal to the diode breakdown voltage in series with a resistor equal to the resistance of the diode after it has broken down and a blocking rectifier to prevent the battery from supplying current to the circuit.

Fig. 3 shows the common emitter characteristics of a typical silicon transistor. As with the diode, the current through the collector junction with the base circuit open is negligible. A locus of points representing 50-milliwatt power dissipation in the transistor is shown together with a typical load line. For example, it would require approximately 120 microamperes flowing into the base of the transistor to obtain a 2-milliampere current into a 10,000-ohm load. This represents not only a current gain but also a power gain because the base input resistance of the transistor is small compared to its output resistance. The other components used in the circuit to be described are familiar to the reader as electrical components, if not as physical components. These are miniaturized resistors, capacitors, and potentiometers.
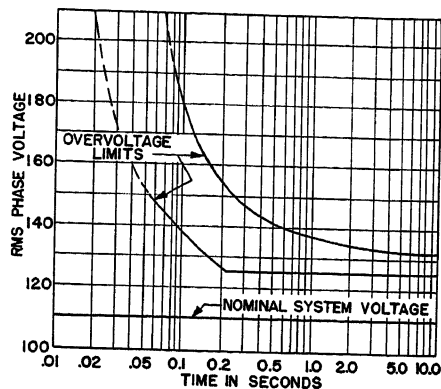
Fig. 1. Operating limits for overvoltage relay[1]

Fig. 3. Silicon transistor characteristics for common emitter connections

Fig. 5. Time-voltage characteristic of relay versus resistance of potentiometer P_2

OPERATION

Fig. 4 shows the circuit of the new overvoltage relay. The a-c relay circuit is illustrated. A similar but more simple circuit may be used on d-c systems. In the diagram, the circuit has been broken down into five basic parts; a voltage-sensing circuit D_1, Z_1, R_1, and P_1; a time-delay circuit D_2, Z_2, R_2, P_2, and C_1; a 2-stage amplifier T_1, and T_2; a power supply D_3, C_2, and TR; and a slave relay SR. Z_1 and Z_2 are Zener diodes, connected in the reverse position and having characteristics such as described in Fig. 2. T_1 and T_2 are silicon transistors having characteristics such as described in Fig. 3. Operation of the circuit is as follows: At normal levels of a-c generator voltage, the direct voltage V_1, which the relay circuit sees, is below the breakdown voltage of Z_1 and the circuit is inactive, with no appreciable current flow anywhere in the circuit. C_1 remains essentially uncharged by the leakage current of Z_1 because of the capacitor discharge circuit R_2D_2. With D_2 connected as it is, the voltage on C_1 cannot exceed the drop across R_2 which is in the order of 0.1 volt. P_1 is set so that at the desired minimum operating voltage of the overvoltage relay Z_1 breaks down and conducts, charging

C_1 through P_2. As C_1 charges, a second Zener diode Z_2 breaks down and the base current I_{b1} starts to flow into T_1. This current when amplified by T_1 becomes the base current of T_2 designated I_{b2}. I_{b2} is amplified by T_2 and becomes the load current I_c. When I_{b1} reaches the critical value, I_c is sufficient to operate the relay.

If at any time before the relay operates the applied voltage returns to normal, the capacitor C_1 discharges rapidly through D_2R_2 and all currents return to zero. This discharge circuit does not affect the time-delay circuit during the charging of C_1 because, as long as the applied voltage exceeds the breakdown voltage of Z_1, the discharge diode D_2 is biased in such a direction that it cannot conduct. R_2 is made small compared to R_1 and P_2 so that the discharge time of C_1 is small compared to its charging time.

ADJUSTMENT OF TIME DELAY

The time delay of the circuit is adjusted by varying the resistance of P_2. Fig. 5 shows the effect of this adjustment. It will be noted that the entire characteristic curve can be shifted over the full adjustment range by merely varying the value of resistance in series with the capacitor. This is in contrast to overvoltage relays in present use which require changes in size and shape of mechanical parts to obtain this adjustment. The characteristic operating curve can be varied over a considerable range by changing the value of P_2 from 5,000 to 20,000 ohms. Further change of resistance in either direction has little effect on the delay time. For adjustment purposes P_2 need only be set to obtain a certain time delay at a single voltage level to assure proper operation over the full voltage range.

Circuit Design

VOLTAGE-SENSING AND TIME DELAY

The development of a suitable sensing and time-delay circuit presents the major design problem. For reasons of co-ordination with other protective devices in the aircraft system, the overvoltage

Fig. 2. Zener diode characteristic

Fig. 4. Circuit diagram of new overvoltage relay
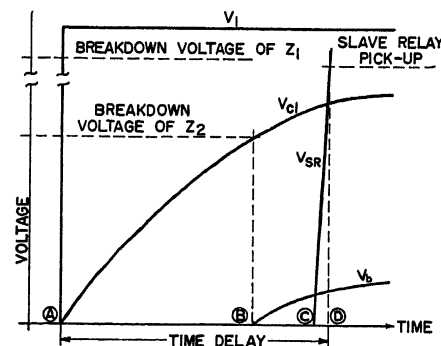
Fig. 6. Circuit voltages versus time on suddenly applied overvoltage
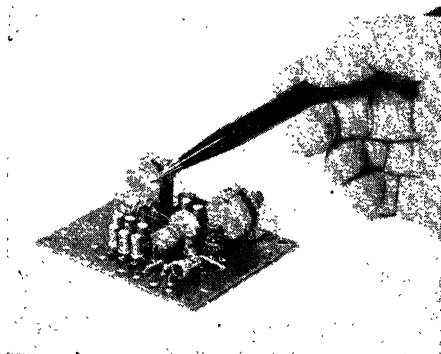
Fig. 7. Components soldered on printed circuit board

Fig. 8. Overvoltage relay package

relay should sense the average of the 3-phase generator output voltages. A direct voltage proportional to the average of the alternating voltage is obtained by using a 3-phase full-wave rectifier made up of six diodes.

The Zener diode Z_1 is used to sense an overvoltage condition. The voltage at which Z_1 breaks down depends somewhat on the ambient temperature but otherwise remains quite fixed. Temperature compensation is accomplished in the resistor combination designated as R_1 in the circuit diagram. This resistance varies with temperature in such a manner that Z_1 breaks down at essentially the same value of line voltage over the entire temperature range of from -55 to $+120$ degrees centigrade.

The time delay obtained at any given voltage is measured by suddenly increasing the voltage from its nominal value to the given overvoltage. If for any reason the capacitor C_1 is partially charged at the moment of overvoltage application, the time delay will, of course, be less than it would be with no preliminary charge on C_1. The ripple on the direct voltage output of the rectifier presents a problem in this type of sensing circuit because the peaks of the ripple voltage start charging C_1 before the average voltage is high enough for relay operation. The ratio of the d-c ripple to the average value of direct voltage is greater by approximately four times than the ratio of the minimum relay operation voltage to the nominal power system voltage. This means that the time delay at a given overvoltage would depend on the nominal voltage level from which the overvoltage was applied because this nominal voltage level would determine the partial charge on C_1.

If a capacitor is used to filter this ripple, true 3-phase average sensing is no longer obtained. A capacitor large enough to reduce the ripple appreciably would also be large enough to hold the

voltage to a level proportional to the highest of the 3-phase voltages rather than the average. It is desirable in some applications to sense the highest phase voltage instead of the average. In that case, a capacitor filter across the rectifier output is desirable and the ripple problem is eliminated.

A full-wave rectifier is used to take advantage of the low ripple resulting from this configuration. Without additional circuitry, the time constant of the time-delay circuit would be the same on charge and discharge of C_1 and, therefore, no particular difficulty would be encountered until the direct voltage ripple was high enough to keep Z_1 broken down and conducting for a longer time than it blocked the charging current. However, this break-even point of charge and discharge is reached at a voltage level below nominal operating voltage, as indicated by the 4 to 1 ratio mentioned in the foregoing.

Additional circuitry in the form of a capacitor discharge circuit D_2R_2 is used to alleviate the aforementioned condition. The operation of this discharge circuit has already been explained as it applies to normal operation. Because of its relatively short time constant, this circuit prevents C_1 from retaining any appreciable charge even when Z_1 is broken down for more than half of each cycle by the d-c ripple peaks. The gain of the amplifier which follows the time-delay circuit is also a factor in the problem of preventing C_1 from being prematurely charged. The gain is made high so that a small change in the voltage being sensed will operate the relay. The use of two stages also makes the individual transistor characteristics noncritical. Any two transistors of the same type may be used in the circuit with good results. The circuitry shown results in a practical circuit with a time-voltage characteristic not appreciably affected by the voltage level from which the overvoltage is applied until this voltage level is very close to the minimum operating voltage of the relay.

TIME DELAY VERSUS VOLTAGE

The time delay obtained at a given voltage depends on the time required to complete the sequence of events necessary for relay operation. Fig. 6 shows this sequence of events in a plot of circuit voltages versus time for a suddenly applied overvoltage. Point A along the abscissa, or time base, is the time at which the overvoltage is applied, i.e., the point when V_1 suddenly rises from zero (or nominal voltage) to some fixed value
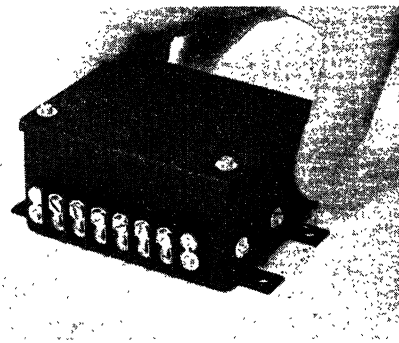
greater than the breakdown voltage of Z_1. The difference between V_1 and the breakdown voltage of Z_1 is applied to the time-delay circuit causing C_1 to charge through P_2. At point B, the charge on C_1 has reached the breakdown voltage of Z_2 and the voltage on the base of the transistor amplifier starts to rise. At point C, the transistor base voltage has reached the threshold value and base current and, therefore, the relay coil current starts to flow causing the voltage across the slave relay V_{SR} to rise. At point D, V_{SR} has reached the pickup value for the relay and it acts to trip open the generator field circuit. The time delay, then, is the elapsed time from point A to point D. An important feature of this design is the fact that no current flows in the slave relay coil until the time delay is almost complete. This means that there is no danger of getting premature operation because of shock or vibration effects on a partially energized relay.

The higher the generator voltage the shorter the time delay before the relay operates. This increase in applied voltage has two effects. It has the effect of increasing the voltage across the resistance-capacitor timing circuit, thus causing the critical value of I_{b1} to be reached sooner; and of increasing the supply voltage on the transistor amplifier, thus causing it to conduct more current for a given value of I_{b1}. This latter effect, however, is small compared to the increased charging rate of C_1. These effects are such that the operating time of the relay varies with applied voltage in a manner which fits the overvoltage relay requirements.

Relay Package

The circuit described herein contains approximately 25 very small parts. It is difficult to mount these small parts in the conventional manner and still take ad-

vantage of their small size. For this reason alone it appears that the use of a printed circuit board is desirable. The other familiar advantages of the printed circuit such as dip soldering and elimination of wiring errors are also realized. Fig. 7 shows the circuit components mounted on a printed circuit board. The transformer and the slave relay are the only parts not mounted on the board. Those two parts are mounted separately.

Fig. 8 shows the complete unit. Six connections are brought out; three for connecting directly to 3-phase line volt-

age and three for the single-pole double-throw relay contacts. The unit is approximately 2 by 3 by 4 inches in size and weighs less than 1 pound.

Conclusions

A new relay circuit is described which has all the desired electrical characteristics of present overvoltage relays while having none of their undesirable features. In addition to meeting the operational requirements, increased reliability and repeatability are realized in

this design. The production problems associated with the manufacture of mechanical time-delay devices have been eliminated and the adjustment and test procedures greatly simplified. The physical package of the new circuit is also described, indicating considerable progress in size and weight reduction, so important to the aircraft designer.

Reference

1. ELECTRICAL POWER, AIRCRAFT CHARACTERISTICS OF. *Military Specification MIL-E-7894*, U. S. Navy Bureau of Aeronautics, Washington, D. C., Amendment 1, Aug. 14, 1952.

Transfer Function and Parameter Evaluation for D-C Servomotors

G. J. THALER
MEMBER AIEE

W. A. STEIN
MEMBER AIEE

IN ANALYZING and synthesizing feedback control systems, the engineer is frequently interested in the numerical value of the parameters associated with the motor-load combination. He probably has manufacturer's data concerning the motor itself, and perhaps similar data are available for many of the load components but, in general, parameter values are not available for the specific motor-load combination and they must be calculated or measured.

With a d-c shunt motor some of the parameters for which numbers might be desired are: armature resistance R, electromotive force (emf) constant K_v, static friction, torque constant, K_t, armature inductance L, and armature inertia J. Some of these parameters are easy to measure and some are not; furthermore, in many cases some of the parameters are negligible.

This paper contains a review of the mathematical equations used for developing the transfer function of the motor and load, an interpretation of these equations in terms of negligibility of certain parameters, the development of measurement techniques, the results of actual tests, and a discussion of the validity of the results. Several unique test methods are presented for determining the moment of inertia of the motor and its load without the necessity of adding tachometers which, in the case of small motors, might greatly alter their operating characteristics.

Mathematical Theory

Fig. 1 shows a schematic diagram of the system to be studied. The motor load is assumed to reflect only additional inertia and viscous friction on the servomotor. The differential equations describing the system are

$$T = K_t i = J \frac{d^2\theta}{dt^2} + f \frac{d\theta}{dt} \quad (1)$$

$$E = iR + L \frac{di}{dt} + K_v \frac{d\theta}{dt} \quad (2)$$

E = applied d-c voltage, volts
i = armature current, amperes
R = armature resistance, ohms
L = armature inductance, henrys
θ = angular position of the motor shaft, radians
T = motor output torque, pound-feet
K_v = motor emf constant, volts per radian per second
K_t = motor torque constant, pound-feet per ampere
J = equivalent inertia of motor, gears and load referred to the motor shaft, slug-feet[2]
f = equivalent viscous friction of motor, gears and load referred to the motor shaft, pound-feet per radian per second

Static friction is taken into account in equation 1 because the torque T is the output torque of the motor, i.e., the developed minus the static torque. The system of units prescribed in the foregoing is somewhat impractical when working with fractional horsepower motors, and the system parameters will be evaluated with the use of more practical units and with appropriate conversion factors applied.

Transforming equations 1 and 2, with the motor assumed initially at rest so that all initial conditions are zero, and manipulating the transformed equations to solve for the motor speed in Laplace form

$$s\Theta_{(s)} = \frac{\dfrac{K_t E}{JL}}{s\left(s^2 + \dfrac{JR + fL}{JL} s + \dfrac{fR + K_v K_t}{JL}\right)} \quad (3)$$

Since there is a quadratic factor in the denominator of equation 3 if a step of direct voltage is applied, the motor speed versus time curve may represent overdamped, critically damped, or underdamped conditions; the exact condition depends on the particular parameter values. Experience shows that most d-c shunt motors of several horsepower or more are appreciably underdamped, those of about 1/4 horsepower may be slightly underdamped, and the still smaller instrument-type d-c servomotors are, in general, overdamped. This paper is concerned primarily with motors of 1/4 horsepower or less. Some of the measurement techniques may be applied to larger motors with satisfactory results, but a few of the techniques give results which require very careful interpretation when the motor is appreciably underdamped.

It has long been customary[1] to treat the d-c servomotor as a single time-constant device. A common assumption is $L = 0$, and $f = 0$, or perhaps $L = 0$, $f \neq 0$. For either assumption, equations 1 and 2
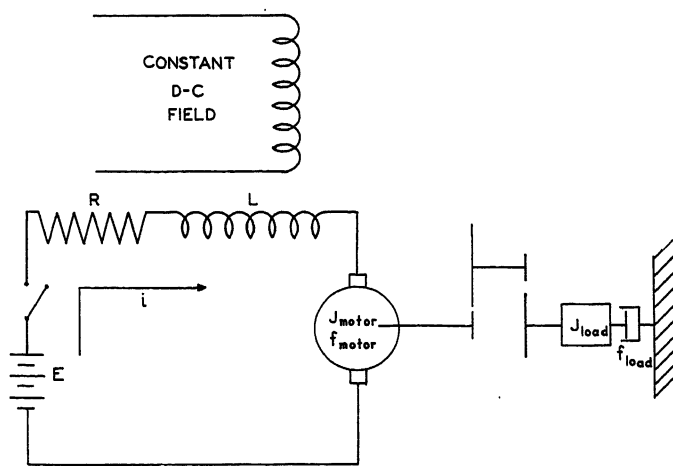
Fig. 1. Schematic diagram of a d-c shunt motor with geared inertia and friction load
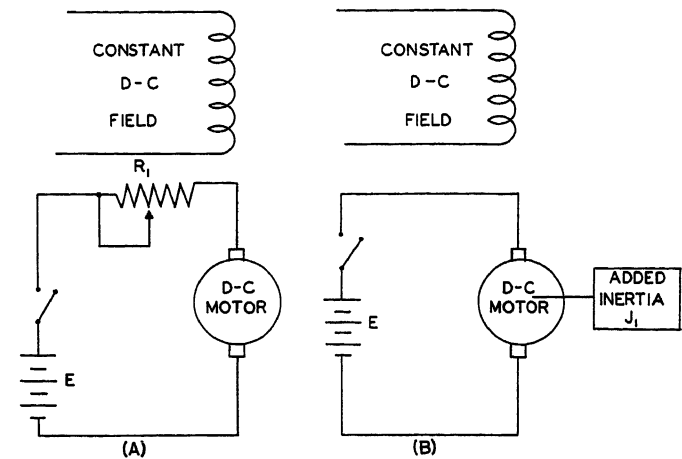
Fig. 3. Servomotor circuits modified to make the mechanical time constant predominate

A—Addition of R_1 in armature circuit
B—Addition of J_1 to motor shaft

must be rewritten and equation 3 would be changed to

$$s\Theta_{(s)} = \frac{\dfrac{K_t E}{JR}}{s\left(s + \dfrac{K_v K_t}{JR}\right)} \quad (4)$$

if the first assumption is used. Using the second assumption

$$s\Theta_{(s)} = \frac{\dfrac{K_t E}{JR}}{s\left(s + \dfrac{fR + K_v K_t}{JR}\right)} \quad (5)$$

In either case the time constant appears as a relatively simple expression, and the parameters involved can be computed.

Equations 4 and 5 give a reasonable approximation of actual performance even though the assumption $L=0$ is not valid.

Returning to equation 3 and factoring the quadratic

$$s\Theta_{(s)} = \frac{\dfrac{K_t E}{JR}}{s\left(s + \dfrac{JR+fL}{2JL} + \sqrt{\left(\dfrac{JR+fL}{2JL}\right)^2 - \dfrac{fR+K_vK_t}{JL}}\right)} \times \frac{1}{\left(s + \dfrac{JR+fL}{2JL} - \right.} \quad (6)$$

$$\sqrt{\left(\dfrac{JR+fL}{2JL}\right)^2 - \dfrac{fR+K_vK_t}{JL}}\right) \quad .$$

If the system is greatly overdamped so that

$$\left(\frac{JR+fL}{2JL}\right)^2 >> \left(\frac{fR+K_vK_t}{JL}\right)$$

then the first factor in the denominator of equation 6 becomes

$$\left(s + \frac{R}{L} + \frac{f}{J}\right)$$

and in general $R/L >> f/J$; therefore, the first factor reduces to

$$\left(s + \frac{R}{L}\right) \quad (7)$$

The second factor in the denominator of equation 6 is determined by the dif-
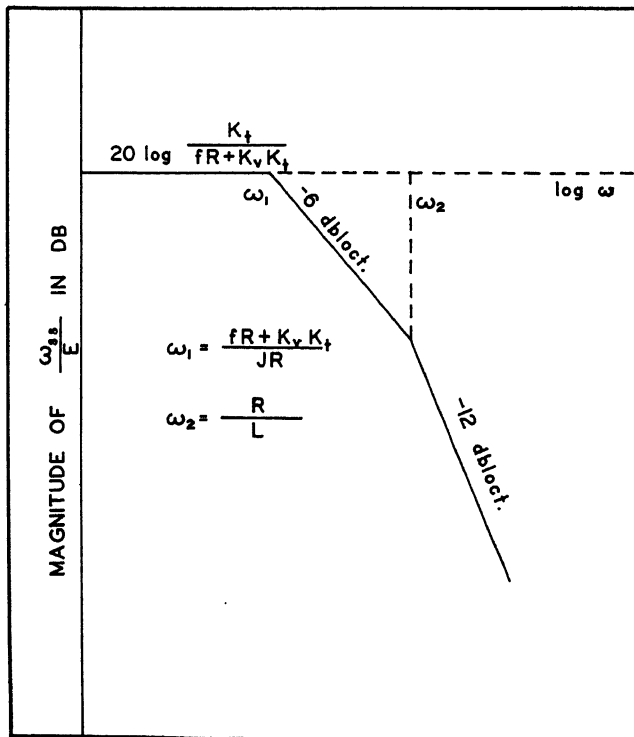
Fig. 2. Asymptotic plot of magnitude ratio versus frequency for motor speed transfer function
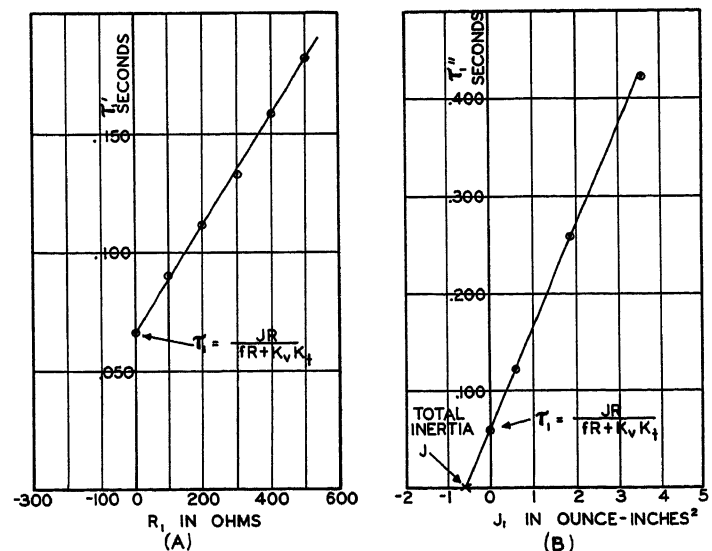
Fig. 4. Straight-line plots used for parameter evaluation
A—Plot of measured time constant versus added armature circuit resistance R_1
B—Plot of measured time constant versus added inertia J_1
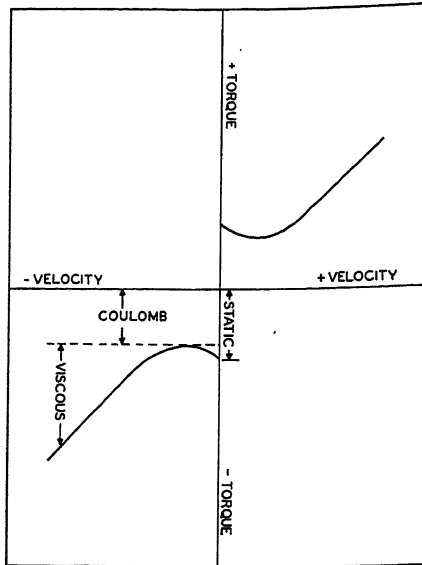
Fig. 5. Types of friction torque

ference between two nearly equal values and, to evaluate this factor, the binomial expansion theorem may be applied to the radical. If differences of higher order are disregarded, the factor reduces to

$$\left(s+\frac{fR+K_vK_t}{JR+fL}\right)$$

Since $JR \gg fL$, this can be simplified to

$$\left(s+\frac{fR+K_vK_t}{JR}\right) \qquad (8)$$

The original equation thus becomes

$$s\Theta_{(s)}=\frac{\dfrac{K_tE}{JL}}{s\left(s+\dfrac{R}{L}\right)\left(s+\dfrac{fR+K_vK_t}{JR}\right)} \qquad (9)$$

It must be recognized that equation 9 is not directly applicable to all d-c motors but is the limit approached when the motor is represented by a greatly overdamped quadratic. This equation is representative of a large class of small d-c servomotors for which the armature resistance is quite large and, furthermore, it may be noted that the limiting time constants are

$$\tau_1 = \frac{JR}{fR+K_vK_t} \text{ seconds (mechanical}$$
$$\text{time constant)} \quad (10)$$

$$\tau_2 = \frac{L}{R} \text{ seconds (electrical time constant)}$$
$$\qquad (11)$$

which are the natural time constants of the mechanical circuit and electric circuit respectively, as would intuitively be expected.

If these two time constants are appreciably different, the smaller has little effect and may be neglected. This is readily seen by rewriting the transformed differential equation as a transfer function expressed in terms of motor speed and plotting on logarithmic coordinates

$$\frac{s\Theta_{(s)}}{E_{(s)}}=\frac{\dfrac{K_t}{fR+K_vK_t}}{\left(s\dfrac{L}{R}+1\right)\left(s\dfrac{JR}{fR+K_vK_t}+1\right)} \qquad (12)$$

The transfer function is given in equation 12, and the asymptotic magnitude plot is shown in Fig. 2. Note that the mechanical time constant is responsible for the lower corner frequency in Fig. 2. The higher corner frequency is set by the electric circuit time constant. If $\omega_2 \gg \omega_1$ the attenuation of ω_2 is large and negligible energy is transferred at the frequency ω_2. Therefore the factor $[s(L/R)+1]$ may be dropped and the transfer function becomes

$$\frac{s\Theta_{(s)}}{E_{(s)}}=\frac{\dfrac{K_t}{fR+K_vK_t}}{\left(s\dfrac{JR}{fR+K_vK_t}+1\right)} \qquad (13)$$

This is exactly the result obtained by assuming that $L=0$; but is not a justification of that assumption, since this result is a limiting expression which is valid only if the motor-load combination is greatly overdamped. However, it does point out a method for measuring the mechanical time constant. If it is possible to make $\omega_2 \gg \omega_1$, then equation 13 is the transfer function of the motor.

It is readily seen that, if R is increased, then ω_1 decreases and ω_2 increases, while if J is increased, ω_1 decreases. Thus $\omega_2 \gg \omega_1$ is obtained either by adding resistance in series with the motor armature or by adding inertia to the motor shaft. For these cases the mechanical time constant becomes respectively

$$\tau_1'=\frac{J(R+R_1)}{f(R+R_1)+K_vK_t} \qquad (14)$$
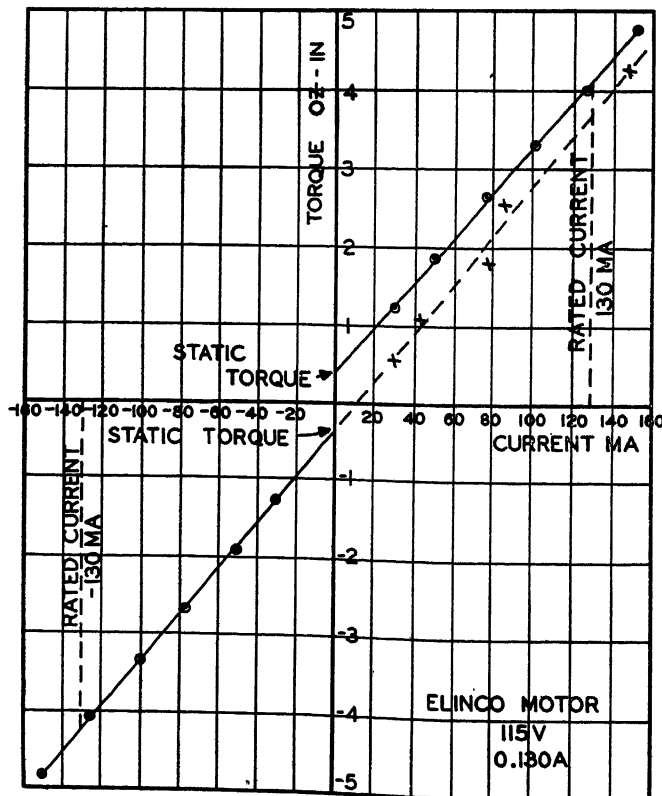
Fig. 6 (left). Stalled torque versus armature current plot for determining K_t and static friction

O—Data from special dynamometer which reads output torque + static friction
X—Data from spring balances which read output torque

Fig. 7 (below). Schematic circuit of speed bridge

Thaler, Stein—Parameter Evaluation for D-C Servomotors
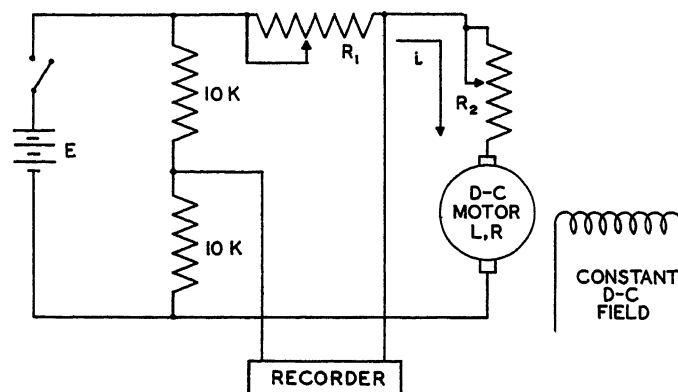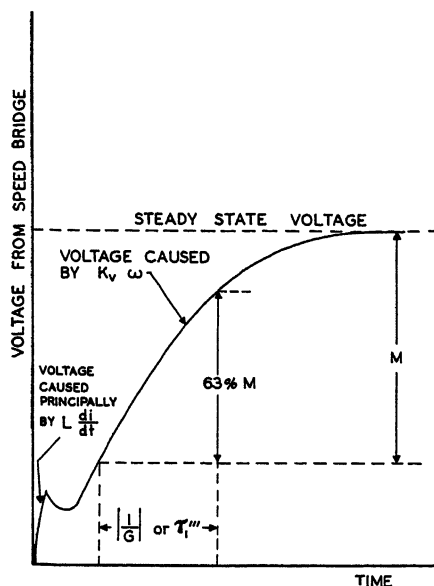
Fig. 8. Brush recorder tape curve showing output of speed bridge

$$\tau_1'' = \frac{(J+J_1)R}{fR+K_vK_t} \tag{15}$$

The time constants of equations 14 and 15 refer to the modified systems, as shown in Figs. 3(A) and (B). The purpose of this analysis, however, is to obtain values of parameters for the unmodified motor-load combination. To accomplish this, a number of successive step-function tests may be made, each for a different value of the modifying parameter. The time constant is measured for each test and plotted against the value of the added parameter R_1 or J_1 as shown in Figs. 4(A) and (B). The curve obtained for the added resistance is a straight line as long as $f(R+R_1)<<K_vK_t$ and this inequality will hold except for very large values of R_1. For added inertia the curve obtained is also a straight line. Extrapolation of either curve to an intercept with the vertical axis provides a value of the mechanical time constant of motor and load

$$\tau_1 = \frac{JR}{fR+K_vK_t} \tag{10}$$

and further extrapolation to an intercept with the horizontal axis provides values for the armature resistance R, as shown in Fig. 4(A), and for the motor-load inertia J, as shown in Fig. 4(B). Note that the addition of resistance to the armature circuit or the addition of inertia to the motor shaft increases the mechanical system time constant and thus slows down the transient response to a step function. This makes the transient response easier to record with an ink-paper recorder like the Brush, and the resulting record may be read with greater accuracy.

Measurement Techniques

Most of the parameters associated with the d-c motor-load combination can be measured by simple, well-known tests. These are reviewed here in outline form to provide a complete coverage. The new techniques developed by the authors are then discussed in detail.

MEASUREMENT OF R

The armature resistance of the motor may be measured by any of the standard bridge methods or at rated current by the voltmeter-ammeter method. The armature should be rotated and measurements made at several positions, since the number of commutator segments short-circuited by the brushes may affect the resistance appreciably.

MEASUREMENT OF K_v

The motor back emf constant may be measured by driving the machine mechanically and measuring the speed and generated emf of the motor.[2] This method, however requires an auxiliary prime mover.

It may also be measured from Kirchoff's law equation

$$E-IR=K_v\omega \tag{16}$$

where

E=the impressed direct voltage
ω=the steady-state motor speed, radians per second

If the load normally includes a tachometer, there is no problem. A very satisfactory technique that can be employed when there is no tachometer is to set a stroboscope on the 60-cycle line frequency and to adjust the magnitude of the impressed voltage E until a revolving disk mounted on the motor shaft appears stationary. A plot of $E-IR$ versus ω will yield a straight line the slope of which is K_v. This method produces a very accurate evaluation of K_v, as it is not necessary to rely on the stroboscope frequency calibration.

FRICTION MEASUREMENTS

There are three types of friction torques involved in the analysis of the d-c motor, namely, static, viscous, and coulomb friction torques.[3] The static torque is the torque that a stalled motor must develop before it can revolve. The viscous torque is a friction torque which is directly proportional to the first power of the speed. The coulomb torque is a friction torque that exists in a rotating motor and is independent of speed. Fig. 5 shows a graphical interpretation of the different types of friction torques.

The static torque can be measured by extrapolating stalled torque versus armature current data, as shown in Fig. 6 and discussed in more detail in the following section. It does not appear in the transfer function of the motor but is a useful parameter and should be evaluated. The coulomb friction is harder to measure. As can be seen from Fig. 5, it is slightly less than the static friction, as the static friction includes the coulomb friction. Engineers are in the habit of using the two interchangeably. The viscous friction can be determined by running the motor under no-load conditions in a special dynamometer[4] designed to read developed torque. In small ball-bearing mounted servomotors it is very small and difficult to measure. In transfer-function evaluation work, the $(fR+K_vK_t)$ term which is of interest can be evaluated as a unit. At first glance it might appear that f could be evaluated from the foregoing expression if R, K_v, and K_t were known, but, in finding a small difference between large terms, the accuracy is limited.

MEASUREMENT OF K_t[5]

The motor-torque constant may be measured by holding the armature at stall, driving current through the armature windings, and measuring the torque; then

$$K_t = \frac{\text{stalled output torque at rated current}}{\text{rated armature current}} \tag{17}$$

The torque may be measured with a dynamometer, prony brake, or a simple system employing a pulley mounted on the motor shaft, some string, and two spring balances. Fig. 6 shows two plots of stalled torque versus armature current for an Elinco motor with field excitation constant at rated value. One set of data was taken with a special dynamometer[6] which measured the total developed stalled torque. The other set was taken by a pair of spring balances which measured the stalled output torque. The static torque as measured by both methods was 0.40 ounce-inch. Thus

$$K_t \text{ by dynamometer} = \frac{4.07-0.40}{0.130} =$$

28.2 ounce-inches per ampere

$$K_t \text{ by spring balance} = \frac{3.72}{0.130} =$$

28.6 ounce-inches per ampere

Although the measurements made with the spring balances are relatively easy, calling for little measuring equipment, it will be noticed that the data points do not

Table I. Evaluated Parameters of an Elinco Shunt Motor

Elinco FD-162 115-Volt D-C Shunt Motor, 1/125 Horsepower, 4,000 Rpm, 0.130-Ampere Armature Current, with a Typical Gear Train and Load

Parameter	Value	Comments
Armature resistance R...	282* ohms.....................	measured by Wheatstone bridge and voltmeter-ammeter method
Emf constant K_v........	0.233* volts per radian per second.....	calculated by equation 16
Static friction...........	0.40* ounce-inch....................	by dynamometer, Fig. 6
	0.40 ounce-inch..................	by spring balances, Fig. 6
Torque constant K_t......	28.2* ounce-inches per ampere.......	by dynamometer, Fig. 6
	28.6 ounce-inches per ampere.......	by spring balance, Fig. 6
Armature inductance L...	0.65* henry........................	d-c field on during all tests, voltmeter-ammeter-wattmeter method at 60 cycles
	0.54 henry........................	General Radio impedance bridge at 1,000 cycles
	0.52 henry........................	one shot resistance-inductance transient on cathode-ray oscilloscope
$(fR+K_vK_t)$ term........	6.76* volts×seconds×ounce-inches radians×amperes	calculated by equation 18
Inertia J................	0.61* ounce-inch²....................	by speed bridge with added R_1, Fig. 9
	0.60 ounce-inch²....................	by added inertia disks with tachometer as part of normal load, Fig. 4(B)
	0.66 ounce-inch²....................	by deceleration test with added inertia disks using armature of d-c motor as speed indicator

* Value used in evaluating transfer function.

line up as well as those obtained with the dynamometer and there might be some hesitancy in drawing the straight line graph.

MEASUREMENT OF L

In measuring the self-inductance of the armature L, the field should be excited at its rated value. Since the self-inductance is composed principally of leakage inductance, the field current and armature current do not have a large effect.[7] The self-inductance was about 10 to 20 per cent higher with the field off in the motors tested.

The self-inductance can be measured by observing the transient growth of current when the armature is excited with a step input of voltage from a source having negligible resistance compared to R. The difficulty experienced in recording this rapid transient and in interpreting the result is discouraging.

Satisfactory results can generally be obtained with a standard voltmeter-wattmeter-ammeter test by energizing the system with a sinusoidal voltage of the lowest frequency available or by using a conventional impedance bridge such as the General Radio impedance bridge-type 650A.

MEASUREMENT OF THE TOTAL INERTIA OF THE MOTOR LOAD WHEN LOAD INCLUDES A TACHOMETER

If the motor load normally includes a tachometer, the circuits employed in either Fig. 3(A) or (B) can be employed. In the first instance, the electrical time constant τ_2 is made negligibly small by the addition of a resistance R_1 in series

with the armature and, in the second case, the same effect is accomplished by the addition of load inertia J_1. The series resistance method is far simpler than the second method, but for the sake of completeness, both methods are presented. The mechanical time constant $\tau_1=(JR)/(fR+K_vK_t)$ can be obtained by the extrapolation of the straight line graphs shown in Figs. 4(A) and (B). Upon applying the final value theorem to equation 13

$$fR+K_vK_t=\frac{EK_t}{\omega_{ss}}$$ (18)

Hence, the expression $(fR+K_vK_t)$ can be evaluated after measuring the ratio of the steady-state speed to the impressed armature voltage. The inertia of the motor and its load is

$$J=\frac{\tau_1 K_t}{R}\left(\frac{E}{\omega_{ss}}\right)$$ (19)

MEASUREMENT OF THE TOTAL INERTIA OF MOTOR AND LOAD USING THE SPEED BRIDGE METHOD

Quite often the motor load does not contain a tachometer, and the addition of one for test purposes would greatly alter the operating characteristics of the motor. A very simple and satisfactory method for measuring the speed-versus-time curve under the aforementioned conditions is the d-c speed bridge illustrated in Fig. 7. The d-c speed bridge has found popular application in circuits employing a d-c motor as an indicator of steady-state speed or for a stabilizing feedback in servomechanisms. As a result, the associated circuitry should be familiar. The application of the bridge for measur-

ing the growth of motor speeds is, however, a novel application and in Appendix I a mathematical justification is presented. The operation of the bridge for time-constant measurement is quite simple. Since the total resistance in the armature circuit is R_1+R_2+R, set R_2 at any arbitrary value and balance the bridge by adjusting R_1. Apply a step of direct voltage to the circuit, as indicated in Fig. 7. The bridge is unbalanced; first, because of the changing current in the armature inductance and, second, because of the back emf, or speed voltage. A typical trace obtained with the brush recorder is shown in Fig. 8. The initial spike is caused by $L(di)/(dt)$ and after this effect has died out, the motor-load time constant can be measured from the balance of the curve.

If the bridge is always balanced so that $R_1=R_2+R$, then a number of time-constant measurements can be made for various values of R_1, the values obtained for τ_1''' may be plotted against R_1, and the resulting curve may be drawn and extrapolated, as shown in Fig. 9. The time constant τ_1''' will approach the following value.

$$\tau_1'''\rightarrow\frac{J(R_1+R_2+R)}{J(R_1+R_2+R)+K_vK_t}$$ (20)

Note that at $R_1=R$, which may be a measured value or may be on the extrapolated curve, the value of τ_1''' is twice the mechanical time constant of the motor-load combination, i.e., $\tau_1'''=2\tau_1$ at $R_1=R$. Equation 19 can then be used to evaluate the inertia of the motor and load.

The speed bridge may also be used to estimate the armature inductance L. If the motor armature is blocked and a step of direct voltage is applied, the speed bridge voltage rises to the value of the applied step voltage and then decays exponentially with a time constant of $L/(R+R_1+R_2)$, from which L may be calculated. On the opening of the switch the transient has a smaller time constant because the energy stored in the inductance discharges through the entire bridge. This method of measuring L is not very accurate unless the recorder has a very high tape speed.

MEASUREMENT OF TOTAL INERTIA BY THE DECELERATION TEST WITH ADDED INERTIA DISKS

The measurement of inertia by the addition of series resistance R_1, with the use of either a tachometer or speed bridge for indicating motor speed, is relatively simple to perform in the laboratory, requiring no special components

Thaler, Stein—Parameter Evaluation for D-C Servomotors
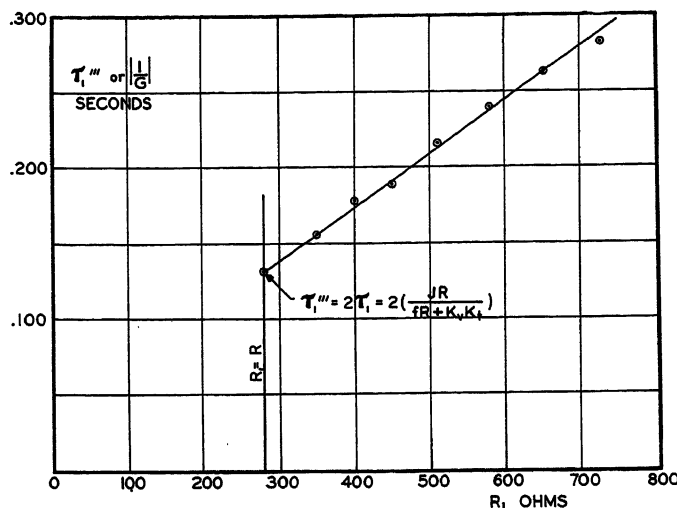
Fig. 9 (left). Plot of measured time constant as read from speed bridge versus added series resistance R_1

$$\tau_i''' = 2\tau_i = 2\left(\frac{JR}{fR+K_vK_t}\right)$$
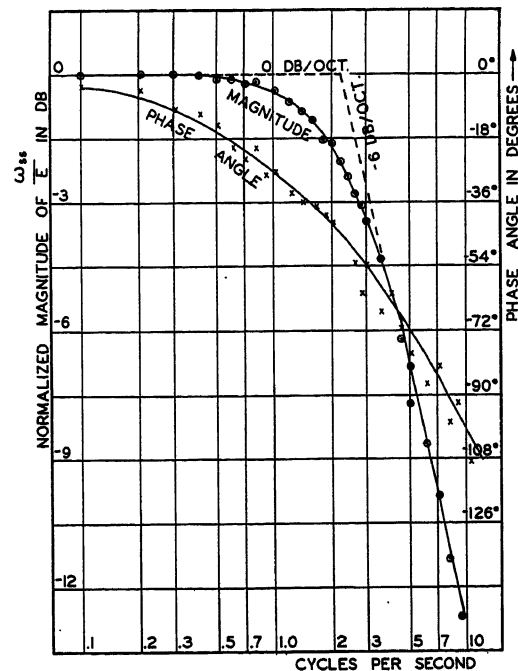
Fig. 10 (right). Bode diagrams for Elinco motor

and it is the recommended method. The method of measuring inertia by the addition of inertia disks not only requires special inertia disks but also presents the problem of how to connect these disks to the motor. Whenever a tachometer is normally part of the motor load, the step-function test with added inertia disks discussed in the section entitled "Measurement of the Total Inertia of the Motor Load when Load Includes a Tachometer" can be used; if a tachometer is not present, a deceleration test with added inertia disks would be simpler than incorporating a speed bridge into the circuit.

The procedure employed for measuring the total inertia by a deceleration test with added inertia disks is as follows.[9] The d-c motor is brought up to an initial speed ω_0 and the armature is disconnected from the supply, allowing the motor to coast to a stop. Since the motor field is left excited, the voltage output of the armature is proportional to speed and can be put on a Brush recorder.

The torque equation of this de-energized motor is

$$J\frac{d^2\theta}{dt^2} + f\frac{d\theta}{dt} + C = 0 \qquad (21)$$

and the instantaneous speed can be determined as

$$\omega = \left(\omega_0 + \frac{C}{f}\right)\epsilon^{-\frac{f}{J}t} - \frac{C}{f} \qquad (22)$$

The time constant of this deceleration curve is $\tau_d = J/f$. By adding series of inertia disks J_1, the time constant will be changed to $\tau_d' = (J+J_1)/f$. A plot of τ_d' versus J_1 will be a straight line, the extrapolation of which to the horizontal axis will yield the desired value of J.

It might be difficult to determine the time constants τ_d' of the different deceleration curves. However, if the initial

speeds ω_0 of all the deceleration tests are the same, it would not be necessary to determine the time constants. The time for the speed to decay to, say, 50 per cent of ω_0 would be proportional to τ_d' and would serve just as well.

Numerical Example

The aforementioned tests were conducted on a series of motors ranging in size from 1/125 to 1/4 horsepower. To present more of an idea of the magnitudes of the parameters involved, the data and calculations for one of these motors are given in Table I. Throughout, curves on

this particular Elinco motor are presented.

The transfer function of the motor and its load becomes

$$\frac{\Theta_{(s)}}{E_{(s)}} = \frac{4.17}{s(0.000152s^2 + 0.066s + 1)} \qquad (23)$$

It will be noted that the parameters given in the foregoing must be converted into the system of units specified with equations 1 and 2 before they can be used in the transfer-function equation to produce equation 23. The damping factor ζ[10] for this motor and load is 2.58 which shows that the motor is considerably overdamped. It is up to the discretion of the engineer whether to use the
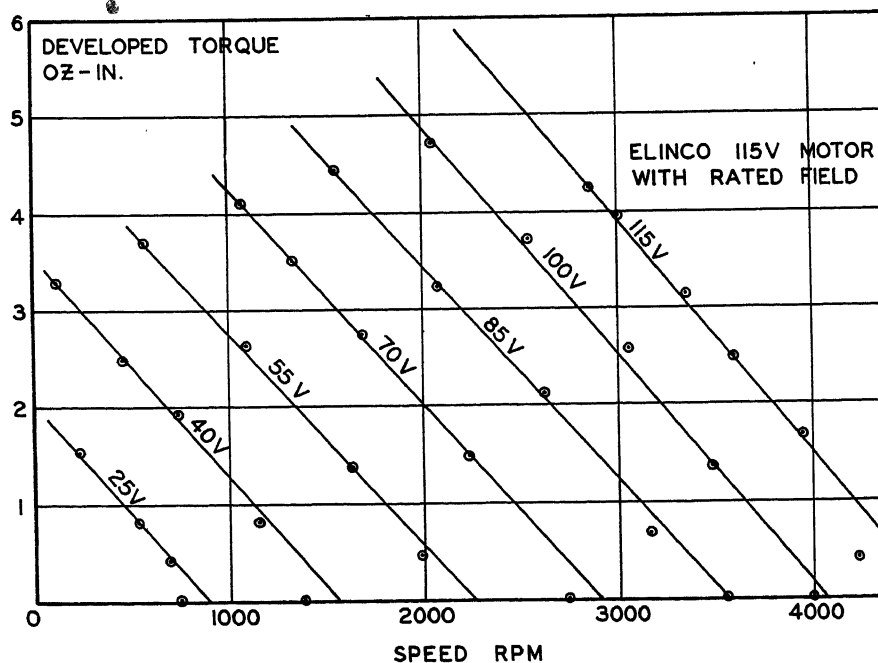
Fig. 11. Torque-speed curves of Elinco motor

transfer function in the form of equation 23 or to simplify it to a single time-constant transfer function

$$\frac{\Theta_{(s)}}{E_{(s)}} = \frac{4.17}{s(0.066s+1)} \qquad (24)$$

In an effort to check out the validity of equation 23, a frequency-response test was performed on this motor. An amplidyne generator with differential fields was used to drive the motor armature. A direct voltage from a battery was put on one differential field and the other was excited from a Hewlett-Packard low-frequency function generator. With this arrangement the motor speed will vary sinusoidally about some fixed speed and, in this way, static friction and backlash in the gearing are eliminated. The input to the armature and the output of a tachometer geared to the motor are recorded on adjacent Brush recorder tapes from which sufficient data are obtained to plot the Bode diagrams of Fig. 10. Unfortunately the phase-angle data are probably only accurate up to about 6 cycles per second, as it is extremely difficult to get angular resolution on the brush recorder tape at high frequencies. The amplidyne had a limited output above about 10 or 12 cycles per second because of the large inductance of its field winding. The intersection of the -6 decibels per octave line with the 0 decibel per octave line occurs at a corner frequency of 2.2 cycles per second which corresponds to a time constant of 0.072 second. According to theory, as the system becomes more overdamped, this value should approach the mechanical time constant of 0.066 second. Although the phase-angle measurements are not closer than about ± 5 degrees above 6 cycles per second, it is quite obvious that the phase shift between output speed and input voltage exceeds 90 degrees and the transfer function has a quadratic term in its denominator. An effort was made to obtain Lissajou figures or two simultaneous waves on the cathode-ray oscilloscope, but the commutator ripple from the amplidyne and d-c motor prevented this.

To present a complete set of data on this motor and to prove that with the exception of static friction or coulomb friction a well-designed d-c servomotor is essentially a linear device, a series of torque-speed curves were measured with the special dynamometer. Total developed torque-versus-speed curves were run for 15-volt increments of voltage and proved to be straight parallel lines, equally displaced from each other, as shown in Fig. 11.

Summary

1. The d-c motor always has a transfer function with a quadratic expression in the denominator. It is up to the discretion of the engineer whether to use the exact transfer function or, under proper conditions, to approximate it by a single time-lag transfer function. When this is possible, the time constant to be used is that of the mechanical system. It might be commented that the small Elinco motor choosen for the numerical example was not a very satisfactory choice since it was greatly damped, $\zeta=2.58$. This particular motor was selected because it was small enough to be driven by the amplidyne generator and also fitted in the special dynamometer. In this way not only was the transfer function evaluated but a very complete analysis and cross check on parameter values could be made. In another case a General Electric 1/4-horsepower shunt motor with $\zeta=0.97$ had a transfer function

$$\frac{\Theta_{(s)}}{E_{(s)}} = \frac{0.775}{s(0.000105s^2+0.0020s+1)}$$

In this instance the exact transfer function would be employed in the analytical or design calculations, the motor being slightly underdamped.

2. A thorough review of standard testing techniques is presented and, in addition, some unique procedures which have proved quite satisfactory. Repeated step-function testing combined with variation of system parameters has given values of mechanical time constant and total system inertia which compare quite favorably with the more elaborate standard methods.

3. A speed bridge has been developed to permit the application of these methods to systems which do not contain a tachometer.

Appendix I. A Mathematical Justification of the Speed Bridge for Measuring Transient Speeds

When R_1 is set equal to R_2+R and the motor is stalled, the circuit of Fig. 7 is in effect a balanced Wheatstone bridge. Whenever the motor is revolving at constant speed, the bridge will be unbalanced by the back emf of the motor. As the motor has a constant field, the back emf will be directly proportional to the speed. No problem is presented when making steady-state measurements; however, when attempting to measure the growth of speed, the unbalance will be equal to the back emf$+L(di)/(dt)$. To measure the growth of speed it is necessary to damp out the masking effect of the $L(di)/(dt)$ terms as quickly as possible.

The differential equations describing the action of the circuit are

$$K_t i = J\frac{d^2\theta}{dt^2} + f\frac{d\theta}{dt} \qquad (25)$$

$$E = ri + L\frac{di}{dt} + K_v\frac{d\theta}{dt} \qquad (26)$$

where $r = R_1+R_2+R$. The time rate of change of current is found to be

$$\frac{di}{dt} = \frac{\frac{E}{JL}(JG+f)}{G-H}\epsilon^{Gt} - \frac{\frac{E}{JL}(JH+f)}{G-H}\epsilon^{Ht} \qquad (27)$$

and the motor speed is

$$\omega_{(t)} = \frac{d\theta}{dt} = \frac{EK_t}{fr+K_vK_t} + \frac{\frac{EK_t}{JL}}{G(G-H)}\epsilon^{Gt} - \frac{\frac{EK_t}{JL}}{H(G-H)}\epsilon^{Ht} \qquad (28)$$

The constants G and H are functions of circuit parameters

$$G = -\frac{Jr+fL}{2JL} + \sqrt{\left(\frac{Jr+fL}{2JL}\right)^2 - \frac{fr+K_vK_t}{JL}} \qquad (29)$$

$$H = -\frac{Jr+fL}{2JL} - \sqrt{\left(\frac{Jr+fL}{2JL}\right)^2 - \frac{fr+K_vK_t}{JL}} \qquad (30)$$

If

$$\left(\frac{Jr+fL}{2JL}\right)^2 > \frac{fr+K_vK_t}{JL} \qquad (31)$$

the speed-versus-time response will be overdamped and the speed approaches its steady-state value exponentially without overshooting. Both G and H will be negative real numbers since r is made sufficiently large to damp out oscillations. The absolute value of H will always be greater than the absolute value of G; and if r is made sufficiently large, then

$$|H| >> |G| \qquad (32)$$

Upon examination of the $(di)/(dt)$ equation, it will be noted that the second term of equation 27, the predominant part, is quickly damped out by the ϵ^{Ht} multiplying factor. In considering the exponential growth of the speed, the second term of equation 28 is larger than the third term. Since G is small, this exponential rise in speed is of long duration in comparison with the $L(di)/(dt)$ voltage.

A plot of speed bridge output voltage versus time as traced out by the brush recorder is shown in Fig. 8. It will be noted that when $r = R_1+R_2+R = 2R_1$ is sufficiently large so that

$$\left(\frac{Jr+fL}{2JL}\right)^2 >> \frac{fr+K_vK_t}{JL} \qquad (33)$$

then

$$\frac{1}{|G|} = \tau_1''' \rightarrow \frac{Jr}{fr+K_vK_t} \qquad (34)$$

and

$$\frac{1}{|H|} \rightarrow \frac{L}{r} \qquad (35)$$

For the aforementioned conditions the masking effect of the $L(di)/(dt)$ is quickly eliminated and the value of $|1/G| = \tau_1'''$ can readily be measured in the manner shown in Fig. 8. For a d-c motor like the small Elinco in the numerical example cited, with a damping ratio $\zeta=2.58$, all eight of the points indicated in Fig. 9 could be read off the speed bridge; whereas with an underdamped motor perhaps only the first five points or so could be read accurately from the speed bridge.

References

1. PRINCIPLES OF SERVOMECHANISMS (book), G. Brown, D. Campbell. John Wiley and Sons, Inc., New York, N. Y., 1948, pp. 129–31.

2. MEASUREMENT OF D-C MACHINE PARAMETER, R. M. Saunders. *Electrical Engineering*, vol. 70, Sept. 1951, pp. 787–92.

3. ELECTRONIC INSTRUMENTS (book), I. Greenwood, J. Holdam, D. MacRae. McGraw-Hill Book Company, Inc., New York, N. Y., 1948, p. 357.

4. A SPECIAL DYNAMOMETER FOR TESTING SMALL MOTORS, S. H. Van Wambeck, W. A. Stein. *Electrical Engineering*, vol. 71, June 1952, pp. 549–51.

5. I. Greenwood, J. Holdam, D. MacRae, *op. cit.*, pp. 453–455.

6. S. H. Van Wambeck, W. A. Stein, *loc. cit.*

7. MEASUREMENT AND CALCULATION OF D-C MOTOR ARMATURE CIRCUIT INDUCTANCE, H.

Snively, P. Robinson. *AIEE Transactions*, vol. 69, pt. II, 1950, pp. 1228–35.

8. THEORY OF SERVOMECHANISMS (book), H. James, N. Nichols, R. Phillips. McGraw-Hill Book Company, Inc., New York, N. Y., 1949, pp. 124–25.

9. EVALUATING THE EFFECT OF NONLINEARITY IN A 2-PHASE SERVOMOTOR, W. Stein, G. Thaler. *AIEE Transactions*, vol. 74, pt. II, Jan. 1955, pp. 518–21.

10. SERVOMECHANISM ANALYSIS (book), G. Thaler, R. Brown. McGraw-Hill Book Company, Inc. New York, N. Y., 1953, pp. 244–45.

———————◆———————

Discussion

T. J. Higgins (University of Wisconsin, Madison, Wis.): The discusser possesses a classified collection of some 3,500 reprints of papers on various phases of feedback control system analysis, synthesis, and components. A review of the 37 papers in the file-folder on d-c servomotors reveals that none of these encompass study of the d-c servomotor as a system component in the degree of thoroughness manifested in the present paper.

The authors are to be congratulated for this excellent and several-sided study of small d-c servomotors. Their review of procedures for obtaining numerical values of pertinent parameters; their substantiation and illustration of these procedures by numerical, experimental example; and their critical study and evaluation of the circumstances determining the equivalent transfer function of a motor of the type studied comprise a paper which merits the careful attention and close reading of all interested in servomechanism theory and practice.

Maximum Overvoltage on Aircraft A-C Generators After Sudden Removal of Load

R. E. KLOKOW
ASSOCIATE MEMBER AIEE

Synopsis: When load is suddenly removed from an a-c generator, the terminal voltage will rise to some maximum value. The magnitude of this overvoltage is a function of several factors such as the amount of load removed, power factor of the load, regulator time delay, exciter response, and the pertinent constants of the a-c generator. This paper presents an analytical study of the influence of these factors on the magnitude of the maximum overvoltage.

THE problem is to determine the effect of the a-c generator parameters on the maximum terminal voltage following sudden removal of load.

Assumptions

The a-c generator is considered as an ideal salient-pole machine, i.e., hysteresis and eddy currents are neglected, the stator magnetomotive forces (mmf) are sinusoidally distributed in space and saturation is not directly included. These are the same assumptions made by Park[1,2] in deriving the direct- and quadrature-axes transient equations of the synchronous machine. Further assumptions

will be made following the ideas presented in papers by Harder and Cheek[3] and Anderson.[4] Both of these papers deal with the problem of calculating the minimum terminal voltage after sudden application of load. These simplifications are: neglect of armature transients, neglect of armature resistance, neglect of speed changes, and neglect of subtransient effects. Rather than include a complete closed-loop system in the analysis, a given voltage as a function of time will be applied to the a-c generator field winding.

Method of Attack

The direct and quadrature components of voltage, current, and flux linkages are used. Kirchoff's equations are set up in operational form. The equations for the direct- and quadrature-axis components of the terminal voltage are solved with the use of the principle of superposition; i.e., the change in voltage is found as determined by opening the stator circuits with the exciter output voltage held constant and then found for the change in exciter voltage. The re-

sulting terminal voltage is found by combining the components. The maximum overvoltage is found by differentiating this expression, equating the results to zero, solving for the time at which the maximum occurs, and substituting in the equation for terminal voltage.

Nomenclature

e_d = direct-axis component of terminal voltage

e_q = quadrature-axis component of terminal voltage

$e_q(0)$ = quadrature-axis component of terminal voltage in steady state before load is removed

e_f = voltage applied to generator field terminals

i_d = direct-axis component of armature current

$i_d(0)$ = direct-axis component of armature current in steady state before load is removed

i_q = quadrature-axis component of armature current

i_f = generator field current

ψ_d = direct-axis armature flux linkages

ψ_q = quadrature-axis armature flux linkages

ψ_f = field winding flux linkages

r_f = resistance of field winding

x_d = direct-axis synchronous reactance

x_d' = direct-axis transient reactance

x_{ad} = reactance of direct-axis armature reactance

x_q = quadrature-axis synchronous reactance

x_{ff} = reactance of field winding

T_{do}' = generator open-circuit time constant
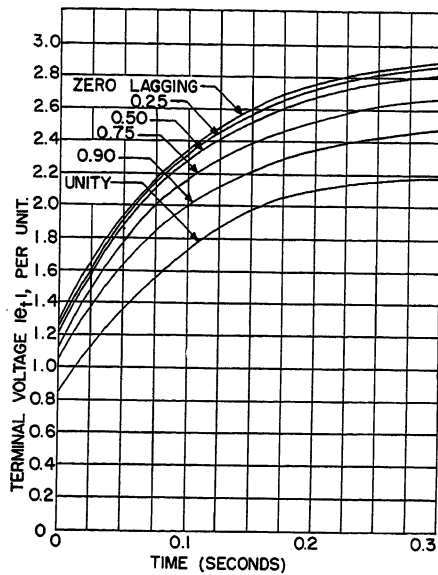
Fig. 1. Terminal voltage after sudden removal of rated load at various power factors, constant exciter output voltage

E = excitation voltage
$p = d/dt$
$\mathcal{L}[f(t)] = F(s)$ = the symbolism for saying that the Laplace Transform of a function of time is a function of the complex frequency s
k = exciter response
t_1 = regulator time delay

Results

An analytical expression for the maximum terminal voltage is given as equation 9 of the Appendix

$$e_{tmax} = e_q(0) + i_d(0)x_d - k \times$$
$$\left[T_{do}' \ln \left(\frac{x_d - x_d'}{k} \frac{i_d(0)}{T_{do}'} + e^{t_1/T_{do}'} \right) - t_1 \right]$$

The effect of the various factors is given explicitly by the foregoing relation, but a brief discussion illustrated with figures follows.

POWER FACTOR

Physically, the power factor of the load determines the position of the armature mmf wave with respect to the position of
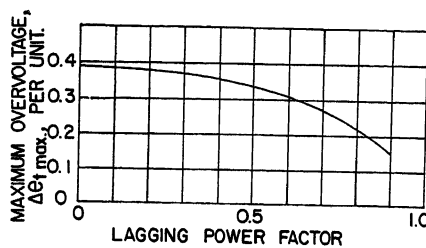
Fig. 2. Maximum overvoltage as a function of power factor of suddenly removed rated load. $x_d = 2.0$ per unit, $x_q = 1.2$ per unit, $x_d' = 0.25$ per unit, $T_{do} = 0.1$ second, $t_1 = 0$, and $k = 100$ per unit per second

the rotor in space. For large inductive reactive loads, the stator mmf is practically in line with and in opposition to the field mmf. This means that more field ampere-turns are required under a low power-factor load in the steady state and, consequently, the terminal voltage without exciter action will rise to a higher value in the new steady-state condition of no load.

The transient in terminal voltage consists of an abrupt change from the former steady-state value followed by an exponential rise, with a time constant determined by the inductance and resistance of the field winding, to the new steady-state condition. Fig. 1 shows the effect of power factor on the envelope of the terminal voltage following sudden removal of load with the exciter voltage held constant. Note that for unity power-factor load, the terminal voltage actually jumps in a negative direction at the instant of switching. Fig. 2 shows the results obtained by considering an exciter response k of 100. As can be seen, the maximum value of the terminal voltage decreases with increase in power factor. The change is not great for power factors in the range of 0 to 0.5, but the value decreases rapidly after 0.5.

Fig. 3. Assumed variation of excitation voltage as a function of time
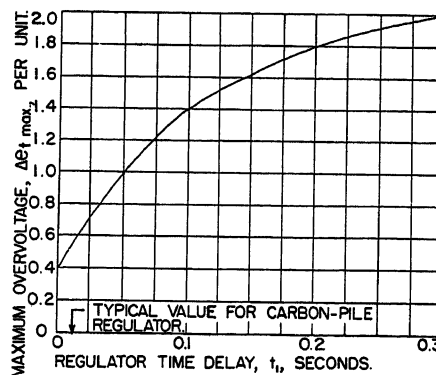
Fig. 4. Maximum overvoltage as a function of regulator time delay. One per-unit zero-power-factor load removed. $x_d = 2.0$ per unit, $x_d' = 0.25$ per unit, $T_{do} = 0.1$ second, and $k = 100$ per unit per second
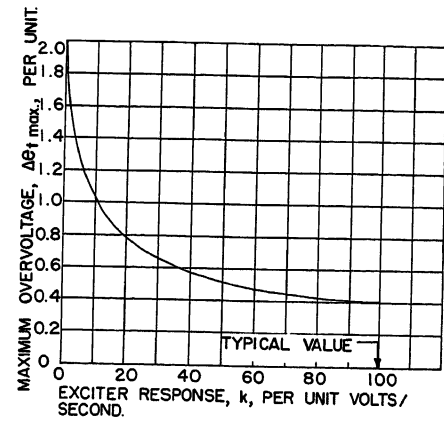
Fig. 5. Maximum overvoltage as a function of exciter response. One per-unit zero-power-factor load removed. $x_d = 2.0$ per unit, $x_d = 0.25$ per unit, $T_{do} = 0.1$ second, and $t_1 = 0$

REGULATOR TIME DELAY t_1 AND EXCITER RESPONSE k

The regulator-exciter combination is assumed to vary, as shown in Fig. 3. The reasons for this simplifying assumption is that it permits the study of the machine without including the complete over-all closed-loop system.

The effect of the regulator time delay on the maximum overvoltage after sudden removal of load is shown in Fig. 4, while the effect of the exciter response is shown in Fig. 5. As would be expected, increasing the regulator time delay makes the maximum overvoltage greater whereas increasing the exciter response decreases the maximum overvoltage.

MACHINE PARAMETERS

Figs. 6 and 7 show the effect of the direct-axis synchronous reactance x_d, the direct-axis transient reactance x_d', and the open-circuit field time constant T_{do}' on the maximum overvoltage.
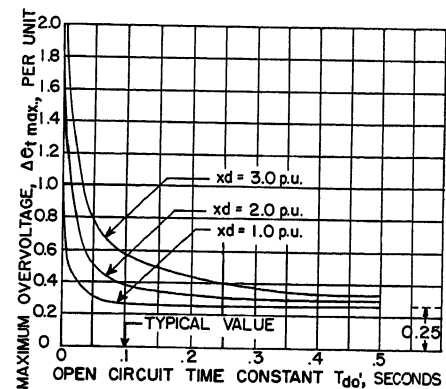
Fig. 6. Maximum overvoltage as a function of a-c generator constants. One per-unit zero-power-factor load removed. $x_d' = 0.25$ per unit, $k = 100$ per unit per second, and $t_1 = 0$

The effect of x_d is relatively large for aircraft machines with time constants under 0.2 second, whereas its effect at larger time constants diminishes. x_d' has a greater effect at the larger time constants than it does at time constants near those of present aircraft machines.

The effect of T_{do}' is very significant in the range of present aircraft machines and its effect levels off at higher values. As can be seen from Figs. 6 and 7, the maximum overvoltage is decreased as both x_d and x_d' are decreased and as T_{do}' is increased.

Effect of Saturation

The previous work neglected the influence of saturation on the results. Saturation effects the results in two ways:

1. The values of the machine parameters are not constant but vary with saturation.

2. The voltage induced in the a-c generator at no load is not a linear function of the field current. For example, in cases where the maximum overvoltage is calculated to be 2.0 per unit, these results would not be correct in a saturated machine. The results do hold more closely for wide speed-range aircraft machines operating at the higher speeds. However, even where the results are not quantitatively correct, they do show qualitatively the effect of the various parameters.

In a particular case saturation can be included by a graphical method given by Rudenberg[5] or with the use of an analogue computer.

Conclusion

The objective of this paper has been achieved, i.e., maximum overvoltage after sudden removal of load is expressed analytically as: a function of the pertinent machine parameters x_d, x_d', and T_{do}'; the power factor of the load removed; the exciter response; and the regulator time delay.

If a particular machine and excitation system were to meet the assumptions made, the results would hold exactly. Every actual system deviates from the simplifying idealizations made here to some degree and the results of this paper must be used with discretion when applied to a particular system. In fact, in a particular case, it would be advisable to use the graphical method of Rudenberg[5] or an analogue computer. However, in preliminary design work of the individual components and in coordinating excitation systems with generators, the broad results given herein will prove useful.
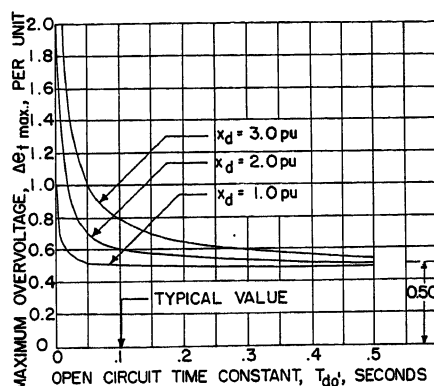
Fig. 7. Maximum overvoltage as a function of a-c generator constants. One per-unit zero-power-factor load removed. $x_d' = 0.50$ per unit, $k = 100$ per unit per second, and $t_1 = 0$

Appendix. Derivation of Expression for Maximum A-C Generator Terminal Voltage

Derivation of Equation

From Park's equations[6] with armature resistance, armature transients and sub-transient effects being neglected

$$e_d = -\psi_q$$

$$e_q = \psi_d$$

$$e_f = p\psi_f + r_f i_f \tag{1}$$

The flux linkages are

$$\psi_d = -x_d i_d + x_{ad} i_f$$

$$\psi_q = -x_q i_q$$

$$\psi_f = -x_{ad} i_d + x_{ff} i_f \tag{2}$$

Eliminating i_f and simplifying

$$e_q = G_{(p)}E - x_d(p)i_d \tag{3}$$

$$e_d = x_q i_q \tag{4}$$

Where

$$G(p) = \frac{1}{1 + T_{do}'p}; \quad x_d(p) = \frac{x_d + x_d'pT_{do}'}{1 + T_{do}p}$$

$$E = \frac{x_{ad}e_f}{r_f}; \quad T_{do}' = \frac{x_{ff}}{r_f}$$
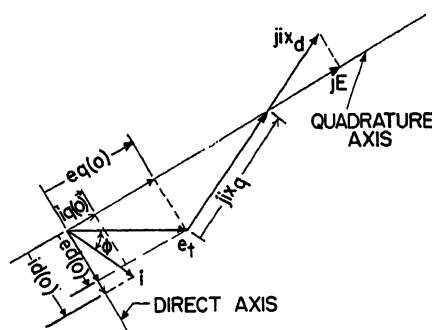
Fig. 8. Steady-state phasor diagram
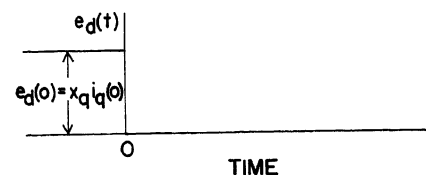
Fig. 9. Direct-axis component of terminal voltage
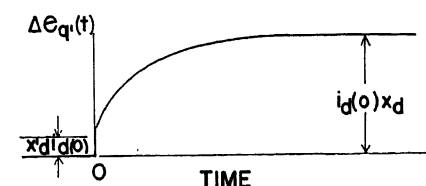
Fig. 10. Component of quadrature-axis voltage caused by opening switch

$$x_d' = x_d - \frac{x_{ad}^2}{x_{ff}}$$

Equations 3 and 4 are the desired relations.

Steady-State Conditions

Prior to removal of load, the machine is operating in the steady-state condition at some given power factor ($\cos \phi$). Conditions are given by the diagram shown in Fig. 8.

Opening Switch and Use of Incremental Voltages

The equations to be solved are put in Laplace transform notation by replacing p with s, since zero initial conditions apply. The equations in the complex frequency domain are

$$E_q(s) = G(s)E(s) - x_d(s)I_d(s)$$

$$E_d(s) = x_q I_q(s)$$

under the conditions that $i_d(t)$ and $i_q(t)$ go from the steady-state values of $i_d(0)$ and $i_q(0)$ respectively to zero at $t = 0$, and $E(t)$ is a given function of time.

To solve these, use will be made of the principle of superposition. First, the change in voltage will be calculated rather than the total voltage. Thus, $\Delta E_d(s)$ and
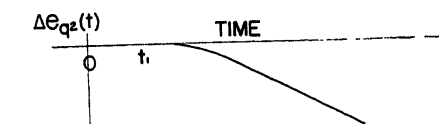
Fig. 11. Component of quadrature-axis voltage caused by change in excitation voltage
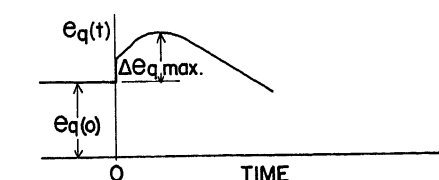
Fig. 12. Total quadrature-axis voltage

$\Delta E_q(s)$ will be the Laplace transforms of the incremental direct- and quadrature axis voltages respectively. In solving for $\Delta E_d(s)$, the effect of opening the switch will be replaced by applying a negative step of current. In solving for $\Delta E_q(s)$, superposition will be applied again. Let

$$\Delta E_q(s) = \Delta E_{q1}(s) + \Delta E_{q2}(s)$$

where

$\Delta E_{q1}(s) = -x_d(s)\ I_d(s) =$ the component of incremental quadrature-axis voltage caused by opening the switch

$\Delta E_{q2}(s) = G(s)\Delta E(s) =$ the component of incremental quadrature-axis voltage caused by change in excitation voltage

The total quadrature- and direct-axis voltage is obtained by adding the initial voltages to these increments. The envelope of terminal voltage is obtained by combining these total components.

Solution for Direct-Axis Component of Terminal Voltage

Equation 4 as a function of time is

$$e_d(t) = x_q i_q(t)$$

Effect of opening switch is given by

$$i_q(t) = i_q(0) - i_q(0)u(t)$$

where

$u(t) = 0$ for $t < 0$

$u(t) = 1$ for $t > 0$

Therefore

$$e_d(t) = x_q[i_q(0) - i_q(0)u(t)] \qquad (5)$$

Fig. 9 shows the direct-axis component of the terminal voltage.

Solution of Quadrature-Axis Components of Terminal Voltage

The component of the incremental quadrature-axis voltage caused by opening switch with excitation voltage constant is

$$\Delta E_{q1}(s) = -x_d(s)I_d(s)$$

Here

$$I_d(s) = \mathcal{L}[-i_d(0)u(t)] = -\frac{i_d(0)}{s}$$

and

$$x_d(s) = \frac{x_d + x_d'T_{do}'(s)}{1 + T_o'(s)}$$

Therefore

$$\Delta E_{q1}(s) = \left(\frac{x_d + x_d'T_{do}'(s)}{1 + T_{do}'(s)}\right)\frac{i_d(0)}{s}$$

Expanding this in a partial fraction expansion and taking the inverse transform yields

$$e_{q1}(t) = i_d(0)[x_d - (x_d - x_d')\epsilon^{-t/T_{do}'}]; \; t > 0 \quad (6)$$

which is graphically shown in Fig. 10.

The component of the incremental quadrature-axis voltage caused by change in excitation voltage is

$$\Delta E_{q2}(s) = G(s)\Delta E(s)$$

Assume

$$\Delta E(t) = -k(t - t_1)u(t - t_1)$$

$$\Delta E(s) = \mathcal{L}e(t) = \frac{-k}{s^2}$$

$$G(s) = \frac{1}{1 + T_{do}'(s)}$$

Therefore

$$\Delta E_{q2}(s) = \left(\frac{1}{1 + T_{do}'(s)}\right)\frac{-k}{s^2}$$

Expanding into partial fractions and taking the inverse transform:

$$\Delta e_{q2}(t) = -k\left[(t - t_1) - T_{do}'\left(1 - \epsilon^{\frac{-(t - t_1)}{T_{do}'}}\right)\right];$$
$$t > t_1 \quad (7)$$

which is shown in Fig. 11.

The total quadrature-axis voltage is

$$e_q(t) = e_q(0) + \Delta e_{q1}(t) + \Delta e_{q2}(t)$$

$$e_q(t) = e_q(0) + i_d(0)(x_d - (x_d - x_d')\epsilon^{-t/T_{do}'}) -$$
$$k\left[(t - t_1) - T_{do}'\left(1 - \epsilon^{\frac{-(t - t_1)}{T_{do}'}}\right)\right] \quad (8)$$

which is illustrated in Fig. 12.

Terminal Voltage

The terminal voltage e_t is determined by using the inverse transformation from the direct- and quadrature-axis components to actual quantities

$$e_t = e_a = e_d \cos \omega t + e_q \sin \omega t + e_0$$

Considering only the magnitude of the terminal voltage and balanced loads

$$|e_t| = \sqrt{e_d{}^2 + e_q{}^2}$$

Since the expression for e_d and e_q are distinct for before and after the switching operation, the terminal voltage is calculated separately for these two periods. For $t < 0$, e_t is given by the steady-state phasor diagram and has been assumed to be 1.0 per unit. For $t > 0$, $e_d = 0$, consequently, $|e_t| = e_q$.

Therefore, the maximum value of $|e_t|$ for $t > 0$ is identical with the maximum value of the quadrature-axis component of the terminal voltage $e_{q\ max}$.

An expression for the maximum terminal voltage is obtained by differentiating equation 8, setting the results equal to zero, and solving for the time of the maximum in equation 8; the following relation for the maximum terminal voltage results

$$e_{t\ max} = e_q(0) + i_d(0)x_d - k \times$$
$$\left[T_{do}' \ln\left(\frac{(x_d - x_d')}{k}\frac{i_d(0)}{T_{do}'} + \epsilon^{t_1/T_{do}'}\right) - t_1\right]$$
$$(9)$$

which is the desired relationship.

References

1. Definition of an Ideal Synchronous Machine and Formula for the Armature Flux Linkages, R. H. Park. *General Electric Review*, Schenectady, N. Y., vol. 31, June 1928, pp. 332-34.

2. Two-Reaction Theory of Synchronous Machines. Generalized Method of Analysis—Part I, R. H. Park. *AIEE Transactions*, vol. 48, July 1929, pp. 716-27.

3. Regulation of A-C Generators With Suddenly Applied Loads, E. L. Harder, R. C. Cheek. *AIEE Transactions (Electrical Engineering)*, vol. 63, June 1944, pp. 310-18.

4. Voltage Variation of Suddenly Loaded Generators, H. C. Anderson, Jr. *General Electric Review*, Schenectady, N. Y., Aug. 1945, pp. 25-23.

5. Transient Performance of Electric Power Systems (book), R. Rudenberg. McGraw-Hill Book Company Inc., New York, N. Y., 1950, pp. 760-64.

6. Synchronous Machines (book), C. Concordia. John Wiley and Sons, Inc., New York, N. Y., 1951, p. 17.

———◆———

Discussion

Morris Flugstad (Boeing Airplane Company, Seattle, Wash.): Mr. Klokow neglects the effects of amortisseur windings in his analysis. We have no data available for such a machine to substantiate experimentally the following conclusions.

Mr. Klokow shows an instantaneous voltage dip upon removing a 1.0 per unit (pu) unity power factor load from an idealized salient-pole a-c generator. Consider the effect of suddenly removing a unity power factor load from a generator which has no amortisseur winding. Flux linkage with the field winding must be the same instantly following a load removal as prior to load removal. Since there is little armature reaction in the direct axis during a unity power factor load, the field instantaneous transient current (d-c component) will be small instantly following load removal. The air-gap flux will thus be nearly the same instantly following load removal as prior to load removal in this ideal machine. (Removing lagging power factor loads will result in an instantaneous increase in air-gap flux of an amount equal to the reduction in field leakage with the resultant transient field current.)

Since the air-gap flux remains the same, the open-circuit voltage of the generator must have an instantaneous voltage rise since the stator leakage reactance drop is zero after load removal.

H. C. Bourne and R. M. Saunders (University of California, Berkeley, Calif.): The author is to be complimented on his solution of a problem of considerable importance in aircraft generator design. His approach is an interesting one and one which would lead to valid results if one were considering the generator alone and not the system as a whole. In his treatment, the author has assumed that the variation of exciter voltage as a function of time is the particular function given in Fig. 3. The assumed variation of exciter voltage has a discontinuous slope at t_1 and t_2. The second discontinuity at t_2 is ignored in the mathematical derivation. This neglect implicity assumes that the time associated with the maximum terminal voltage is less than t_2. For large

enough K or $E_{residual}$, the assumption is not true and such a mathematical boundary should be defined. In addition, it would have been exceedingly interesting to explore the effects of zero and negative values of $E_{residual}$, the negative values corresponding to exciter forcing. The zero value of $E_{residual}$ and large K might well correspond to magnetic amplifier operation where the time of response is exceedingly rapid, of the order of one-half cycle or so. In the interests of completeness, the author should extend the treatment to cover these cases.

As pointed out by the author, one must consider the saturation effects since they in general will limit the peak overshoot rather than the linear reactances as assumed.

An exceedingly simple but not unrealistic method to take into account machine saturation is to put an upper bound on the maximum voltage overshoot. In fact, indications are that operation of machines at higher flux densities is sound from both weight and overshoot consideration if the problem of harmonic distortion can be overcome by better design or relaxation of specifications.

R. E. Klokow: Mr. Flugstad is correct in saying there cannot be an instantaneous drop in terminal voltage after removal of unity power factor load from a machine having a damper bar circuit in the quadrature axis, but this statement is incorrect for an ideal machine without damper bars. This can be demonstrated as follows.

From equation 3, the quadrature axis terminal voltage the instant after removal of load is

$$e_q(0+) = e_q(0-) + i_d(0-)x_d'$$

where the plus and minus signs mean the instant after or before $t=0$, respectively. The direct-axis voltage the instant after is zero.

$$e_d(0+) = 0$$

The terminal voltage is

$$e_t(0+) = \sqrt{e_d{}^2(0+) + e_q(0+)^2} = e_q(0+)$$

Now it is indeed possible that at unity power factor $e_q(0+)$ is less than $e_t(0-) = 1.0$. Consequently the terminal voltage immediately after load is removed is less than 1.0. For example, let $x_q = 1.0$ pu and $x_d' = 0.25$ pu, then

$$e_q(0-) = 0.7$$

and

$$i_d(0-) = 0.7$$

$$e_q(0+) = 0.875$$

Therefore, in this case terminal voltage drops instantaneously from 1.0 pu to 0.875 pu.

Physically, the point is that although the quadrature-axis terminal voltage (generated by the direct-axis flux) rises from 0.7 to 0.875, the direct-axis voltage (generated by the quadrature-axis flux) drops from 0.7 to 0. The flux in the direct axis is maintained by the induced field current, but the flux in the quadrature axis collapses as soon as the armature current goes to zero.

The replies to comments of Professors Bourne and Saunders are a necessary addition to the paper. The author agrees that mathematical boundaries should be placed on the results. Equation 9 of the Appendix holds for $t_1 < t_{max} < t_2$ where t_{max} is the time at which the maximum voltage occurs. This condition was met for the curves given in the paper for the case of small residual. To demonstrate that large values of k do not invalidate the results, a numerical example follows:

Let

$$x_d = 2.0, \ x_d' = 0.25, \ T_{do}' = 0.1, \ t_1 = 0.1, \ E_{res} = 0$$

with

$$k = 10, \ t_{max} = 0.09, \ t_2 = 0.3$$

while for

$$k = 100, \ t_{max} = 0.004, \ t_2 = 0.03$$

It appears that as k increases, t_{max} decreases such that the $t_{max} < t_2$. In a practical case where the residual is small compared to the excitation voltage under load, the results given in the paper hold. However, for mathematical completeness, the case of $t_{max} > t_2$ will be discussed. For this case equation 8 of the Appendix would be

$$e_q(t) = e_q(0) + i_d(0)[x_d - (x_d - x_d')e^{-t_1/T_{do}'}]$$

$$-k\left\{(t-t_1) - T_{do}'\left[1 - e^{-\frac{(t-t_1)}{T_{do}'}}\right]\right\}$$

$$+k\left\{(t-t_2) - T_{do}'\left[1 - e^{-\frac{(t-t_2)}{T_{do}'}}\right]\right\}$$

for $t > t_2$

An equation for the maximum terminal voltage can be derived in exactly the same manner as equation 9 of the Appendix.

Other functions of excitation voltage can be assumed better to meet a particular type of regulator system. The author has tried step functions and exponentials, but the results are too cumbersome to be of value in this discussion. It is the opinion of the author that "exciter forcing" would be of value in decreasing the time required to return to normal voltage after the maximum has been reached, but that the exciter response (k) is the more important factor in limiting overvoltage.

The suggestion of Professors Bourne and Saunders of considering saturation as an upper bound on the maximum voltage overshoot is a most effective method of including saturation since it does not invalidate the superposition used in the analysis and yet prevents erroneous answers. Generator designers at the author's company have taken advantage of this fact to the extent that on some constant speed machines, the maximum voltage is limited to 140 per cent. Saturation introduced no detrimental effect on the harmonics content.

Design of High-Precision Synchros and Resolvers

R. A. HEARTZ
ASSOCIATE MEMBER AIEE

R. M. SAUNDERS
MEMBER AIEE

Synopsis: In the design of high-precision resolvers and synchros there are many sources of inaccuracies, such as harmonics generated by winding distributions, slot openings, and mechanical tolerances. This paper develops the equations for output voltage in terms of winding and air-gap factors, gives a method of analysis which is adaptable to digital techniques, and presents a general discussion of the sources of error and a means for minimizing these errors.

THE ever-increasing emphasis upon accuracy in the field of control systems has called for higher precision transducers to translate mechanical position into electrical signals and to perform other computations. Synchros and resolvers are two devices widely used in this capacity whose similarity is such that they may be discussed together. Each of these has a rotor whose single winding generates a sinusoidal space distribution of flux density. In the resolver the windings on the stator are effectively spaced 90 degrees apart, whereas in the synchro there are three windings effectively 120 degrees apart. Fig. 1 shows the electrical and mechanical arrangements of the coils and the desired or ideal relationships between the supply voltage, the
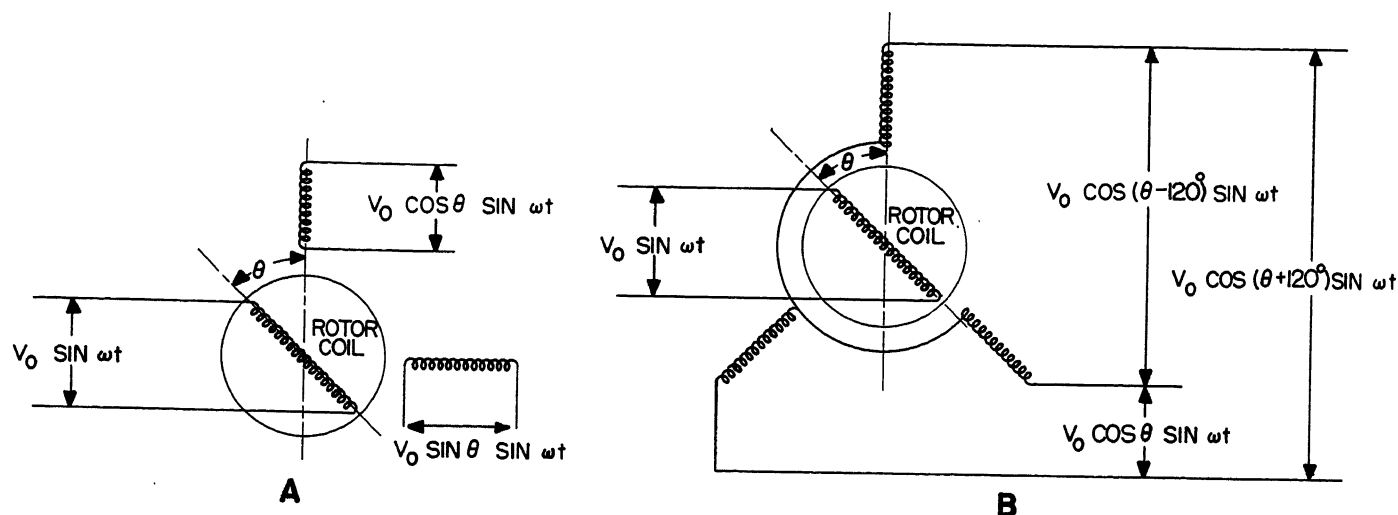
Fig. 1. Voltage relationships

A—For a resolver
B—For a synchro

displacement angle, and the output voltage. The output voltages are induced in the stator by the time rate of change of the supply voltage. Thus synchros and resolvers are essentially transformers whose mutual coupling is variable.

The history of the synchro is much the same as that of any other high-precision device. It was originally developed for applications where accuracy was not of primary importance. A synchro generator was used to drive directly a follow-up synchro receiver. In 1934, an early paper[1] gives an excellent analysis of this system. The following decade, 1940 to 1950, saw the development of the synchro as an informational rather than as a power transmitting device. Early units were designed to have a linear relationship between output voltage and angular displacement over a range of 60 to 90 degrees, but current practice in most cases is to attempt to achieve a sinusoidal voltage-displacement relationship. To obtain higher accuracies synchros are now used in closed-loop feedback control systems in a transmitter-control transformer combination. This application differs from the earlier use in that the synchro transmitter serves to sense the mechanical position, transmitting this information to the control transformer where a comparison is made with the reference input to result in a difference or error signal. In closed-loop applications synchros operate at low-power levels and have the primary function of transmitting information. An analysis of the electrical accuracy of this type system was presented by Chestnut[2] in 1945. A definitive analysis of the errors involved in a salient pole-type synchro generator-control transformer system was presented by Kronacher[3] in 1950. The accuracies of 1 per cent (%) which were common in the last decade, have given way to accuracy demands of better than 0.1% which are necessary in many of today's systems. Thus, this paper is concerned with the design of synchros and resolvers whose relationship between shaft position and output voltage will not deviate more than 0.1 % from a sinusoid. The authors have found it necessary to limit the scope of the paper to factors concerning the accuracy of the synchro or resolver unit rather than considering the complete synchro or resolver system.

For a practical case the relationship between the output voltage induced in a stator winding and the shaft position can be represented by a Fourier series as follows. (It is assumed throughout the paper that the velocity is zero.)

$$e_0(\theta, t) = k \sum_{n=1}^{\infty} a_n \cos (n\theta + \lambda_n) \cos (\omega t + \psi)$$

(1)

The harmonic amplitudes of this series are directly related to the errors of the unit. The space phase position of the fundamental in the foregoing series is also a source of error (i.e., the voltages induced in the stator windings are not 90 or 120 degrees apart in space phase position). Time harmonics and time phase position can also be a source of error. However, the errors which are functions of time are generally not directly related to the accuracy of the unit and are usually of minor concern to the designer.

The purpose of this paper is to isolate the various sources of error and to present design techniques for minimizing these errors. To accomplish this goal, the general equations for the output voltage in terms of winding and air-gap factors are developed in the first part of the paper. The methods of analysis and the final equations have been specifically designed so that they are adaptable to digital computer techniques. A general discussion of the sources of error and means for minimizing errors are given in the second part of the paper. Although the paper is devoted primarily to factors which affect the precision of resolvers and synchros, the method of analysis and many of the conclusions of the analysis apply directly to most high-precision rotating magnetic devices such as induction potentiometers, velocity generators, etc. It is the authors' hope that others will be stimulated to contribute to the subject through discussion of this paper or by subsequent papers, because this field of design involves practical experiences and opinions which cannot always be reduced to algebraic equations.

Nomenclature

The subscripts r and s refer to rotor and stator respectively.

a_n, b_n = harmonic coefficients of general Fourier series
$\bar{a}_n$ = harmonic reduction factor, normalized
$B(x, y, t)$ = air gap flux density, gausses
B_{nm}, B_{nm}' = harmonic coefficients of air gap flux density Fourier series, gausses
$\bar{B}_{nm} = \sqrt{B_{nm}^2 + B_{nm}'^2}$
$e(\theta, t)$ = stator output phase voltage as a function of rotor position, volts

Heartz, Saunders—Design of High-Precision Synchros and Resolvers

Fig. 2. Rotor coil parameters

$$\bar{a}_{nr} = \sqrt{\frac{a_{nr}^2 + b_{nr}^2}{a_{1r}^2 + b_{1r}^2}}$$

$$\bar{\lambda}_{nr} = \tan^{-1}\frac{b_{nr}}{a_{nr}}$$

$$a_{nr} = \frac{\sin n\sigma/2}{n\sigma/2}\sum_{j=1}^{qr} N_{rj}\frac{\sin n\frac{\beta_{rj}}{2}}{n}\cos n\alpha_{rj}$$

$$b_{nr} = \frac{-\sin n\sigma/2}{n\sigma/2}\sum_{j=1}^{qr} N_{rj}\frac{\sin n\frac{\beta_{rj}}{2}}{n}\cos n\alpha_{rj}$$

$$k_r(t) = 1.13\sqrt{a_{1r}^2 + b_{1r}^2}\,I_r\sin\omega t$$

Equation 3 of reference 4 gives the general relationship for the flux density in the air gap

$$B(x, y, t) = f(x, t)P(x, y) \tag{3}$$

The permeance function, $P(x,y)$, is given by equation 4 of the previous paper

$$P(x, y) = \sum_{\mu=-\infty}^{\infty}\sum_{\nu=-\infty}^{\infty} P_{\mu\nu}\cos(\mu x + \nu y) + P_{\mu\nu}'\sin(\mu x + \nu y) \tag{4}$$

In that paper methods were outlined for determining the coefficients $P(\mu,\nu)$ and $P'(\mu,\nu)$ for any given configuration, thus it will be assumed in this paper that these terms are known.

Performing the multiplication indicated in equation 3 yields an air-gap flux density expression which can be expressed in the following form

$$B(y, \theta, t) = k_B(t)\sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\bar{B}_{nm}\cos(my + n\theta + \lambda_{nmB}) \tag{5}$$

where θ is the relative position of the rotor reference ($x=0$) with respect to the stator reference ($y=0$). Thus $x=y+\theta$ and $\bar{B}_{nm}$ are flux density amplitudes resulting from the product of the air gap mmf and permeance expressions.

The voltage induced in the jth stator coil whose center position is α_{sj} and whose pitch is β_{sj}, as shown in Fig. 3, is

$$e_j(\theta, t) = lr_g N_{sj}\frac{d}{dt}\int_{\alpha_{sj}-\frac{\beta_{sj}}{2}}^{\alpha_{sj}+\frac{\beta_{sj}}{2}} B(y, \theta, t)dy\times10^{-8}$$

The voltage induced in a stator phase belt will be the summation of the individual coil voltages. Performing this integration and summing the coil voltages yields

$$e(\theta, t) = k_s\frac{d}{dt}k_B(t)\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\bar{B}_{nm}\bar{a}_{ms}\times$$
$$\sin(n\theta + \lambda_{nmB} + \lambda_{nms}) \tag{6}$$
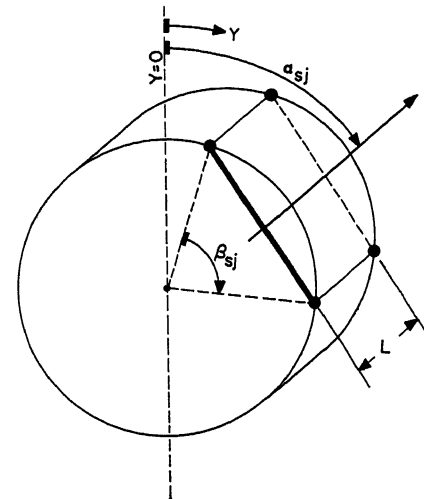
Fig. 3. Stator coil parameters

where $\bar{a}_{ms}$ is the normalized stator harmonic reduction factor and has the same defining expressions as the normalized rotor harmonic reduction factor $\bar{a}_{mr}$ equation 2, and

$$k_s = \frac{lr_g\times10^{-8}}{2}\sqrt{a_{1s}^2 + b_{1s}^2}$$

Equation 6 is far too complex to treat as a general case. However, for a specific case, the equation simplifies so that results may be readily obtained. This is particularly true if the analysis is made with the aid of a digital computer.

The specific situation of a uniform air gap is a case in point in that equation 6 simplifies to

$$e(\theta, t) = \frac{lr_g}{2l_g}\times10^{-8}\frac{d[k_r(t)]}{dt}\sum_{n=1}^{\infty}\bar{a}_{nr}\bar{a}_{ns}$$
$$\cos(n\theta + \lambda_{nr} + \lambda_{ns}) \tag{6A}$$

Equation 6(A) is the starting point in any preliminary design, for this equation will show the magnitude of output voltage harmonics generated when air-gap effects are neglected. Thus, equation 6(A) shows directly the effectiveness of the stator and rotor harmonic reduction factors. The addition of the air-gap effects will in general increase the output harmonic content. Thus, the analysis of the unit with uniform air gap establishes conditions of necessity but not sufficiency.

As a second example, consider the case of a rotor with an off-center shaft but with an otherwise uniform air gap so that the expression for the air-gap permeance is a function of x alone. As a consequence of the uniformity of the rotor and stator surface, it may be assumed that all permeance harmonics are negligible when compared to the fundamental. Thus

Development of General Equation for Output Voltage

In a previous paper[4] the authors have developed general expressions for the airgap mmf and permeance. Equation 2(A) of that paper gives the general expression for the mmf developed by a group of coils, one of which is shown in Fig. 2. Stator coil parameters are shown in Fig. 3. For convenience in future analysis this equation is restated here in terms of normalized reduction factors and phase positions of the harmonics:

$$f(x, t) = k_r(t)\sum_{n=1}^{\infty}\bar{a}_{nr}\cos(nx + \lambda_{nr}) \tag{2}$$

where

$f(x, t)$ = magnetomotive force (mmf), gilberts
I_r = rms value of rotor current, amperes
j = integer denoting single coil
l = effective stack length, centimeters (cm)
l_g = minimum air gap length, cm
M = figure of merit
μ, ν, n, m = order of harmonic, an integer
N = turns per coil
$P(x, y)$ = air gap permeance function
$P_0,\ P_{\mu\nu},\ P_{\mu\nu}'$ = harmonic coefficients of permeance function
r_g = radius at the air gap, cm
S = number of slots
t = time, seconds
x = angular distance around rotor, degrees
y = angular distance around stator, degrees
α = coil distribution angle, degrees
β = coil pitch angle, degrees
θ = angle between stator reference and rotor reference, degrees
ω = supply frequency, radians per second
λ_n = angle associated with harmonic reduction factor, degrees
σ = skew angle, radians

$$P(x, y) = (P_0 + P_1 \cos x)$$

and

$$B(y, \theta, t) = k_r(t) \sum_{n=1}^{\infty} P_0 \bar{a}_{nr} \cos(nx + \bar{\lambda}_{nr}) + \frac{P_1 \bar{a}_{nr}}{2} \{\cos[(n+1)x + \bar{\lambda}_{nr}] + \cos[(n-1)x + \bar{\lambda}_{nr}]\}$$

For the special case where there is a large fifth harmonic in the mmf and no others such as

$$f(x, t) = k_r(t)[\bar{a}_{1r} \cos x + \bar{a}_{5r} \cos(5x + \bar{\lambda}_{5r})]$$

the output voltage will be

$$e(\theta, t) = k_r(t)k_s \Big\{ P_0[\bar{a}_{1r}\bar{a}_{1s} \cos(\theta + \bar{\lambda}_{1s}) + \bar{a}_{5r}\bar{a}_{5s} \cos(5\theta + \bar{\lambda}_{5r} + \bar{\lambda}_{5s})] + \frac{P_1}{2}[\bar{a}_{1r}\bar{a}_{2s} \cos(2\theta + \bar{\lambda}_{1r} + \bar{\lambda}_{2s})] + \frac{P_1}{2}[\bar{a}_{5r}\bar{a}_{6s} \cos(6\theta + \bar{\lambda}_{5r} + \bar{\lambda}_{6s})] + \frac{P_1}{2}[\bar{a}_{5r}\bar{a}_{4s} \cos(4\theta + \bar{\lambda}_{5r} + \bar{\lambda}_{4s})] \Big\}$$

Thus it is seen that new harmonics are generated by air-gap effects.

Although the foregoing example is a simple one, it is seen that the steps used in driving the output voltage expression are such that a more complex problem can be handled by ordinary or digital computer techniques. The general steps in the analysis are:

1. Determine the expression for the air-gap mmf and the air-gap permeance function.

2. Determine the air-gap flux density by taking the product of these two functions and then arrange this expression into the form of equation 5.

3. Multiply each air-gap flux density expression by the stator harmonic reduction factor which corresponds to the coefficient of y.

4. Sum all terms, with proper regard to phase positions, which have like coefficients of θ.

5. Multiply the final expression by the stator constant, k_s.

This will yield an expression for the output voltage which is in the form of equation 1.

Discussion of Errors

As stated previously, the accuracy of a resolver is directly proportional to the harmonic amplitudes of the Fourier series for the stator output voltage as a function of rotor position. This Fourier series is developed in the first part of the paper. The sources of space harmonics are attributable to several factors. The winding

itself is the prime source of space harmonics in that the mmf on the rotor surface generated by a coil will have a rectangular wave form. Proper choice of coil pitch, distribution, and number of turns per coil is used to minimize these harmonics. The fact that the coil distribution must vary in discrete steps equal to the number of slots is a source of a large harmonic whose frequency is equal to the number of slots. Proper choice of the number of rotor slots and stator slots will minimize this harmonic. Slot openings which cause a variation in the air-gap permeance will also produce output space harmonics. To remove these harmonics, the rotor or stator is skewed. Skewing will also help remove the winding step harmonics. Finally, the minimum basic tolerances of the manufacturing process are a prime source of harmonics in the stator output voltage. This source can be attenuated only by the selection of a winding design which offers the maximum insensitivity to manufacturing errors.

Winding Harmonic Reduction Factors

The winding harmonic reduction factor $\bar{a}_{nr}$ or $\bar{a}_{ns}$ can be expressed as follows

$$k\bar{a}_n = \frac{1}{n} \sum_j N_j \sin n\frac{\beta_j}{2} \cos n\alpha_j \qquad (7)$$

Equation 7 assumes that the winding is symmetrical about the direct axis so that any quadrature axis flux generated by a set of coils is cancelled. Thus, the summation of the coil distribution angles must be equal to zero. This, as will be evident in the following discussion, is a condition which must be met in high precision resolvers or synchros.

Let $\bar{a}_{jn}$ be the amplitude of the nth harmonic generated by the jth coil. A series of simultaneous equations can be set up equating the harmonics generated by each coil to zero and the fundamental to a constant. These equations are

$$\bar{a}_{11} + \bar{a}_{21} + \bar{a}_{31} \ldots \bar{a}_{j1} = k_1 \text{ (fundamental)}$$
$$\bar{a}_{13} + \bar{a}_{23} + \bar{a}_{33} \ldots \bar{a}_{j3} = 0 \text{ (3rd harmonic)}$$
$$\bar{a}_{15} + \bar{a}_{25} + \bar{a}_{35} \ldots \bar{a}_{j5} = 0 \text{ (5th harmonic)}$$
$$\vdots$$
$$\bar{a}_{1n} + \bar{a}_{2n} + \bar{a}_{3n} \ldots \bar{a}_{jn} = 0 \text{ (nth harmonic) (8)}$$

The simultaneous solution of these equations will give a winding design which will provide maximum suppression of harmonics. The number of harmonics which can be suppressed by a proper choice of $\bar{a}_{jn}$ is one less than the number of coils of the winding.

Theoretically, either the pitch factor, $\sin n\beta_j/2$, the distribution factor, $\cos n\alpha_j$ or the number of turns per coils, of equation 7 could be varied to effect a solution of equation 8. However, there are a number of practical limitations on varying these parameters. In the first place, a solution which involves negative turns is not physically realizable. Second, the pitch factors and distribution factors can be varied only in large steps equal to the slot pitch, $360/S$ degrees where $S =$ the number of slots. Further, the winding must be symmetrical and the distribution angle should not exceed ± 10 degrees. For distribution angles larger than 10 degrees, a large part of the flux generated is quadrature. Although the winding is designed symmetrically to cancel the quadrature component of flux, the effectiveness of this quadrature cancellation is limited by mechanical tolerances. Thus, a winding which has large quadrature fluxes requires tight tolerances in mechanical construction or, if these tight tolerances cannot be maintained, there will be a high rejection rate of the finished units. Therefore the distributed winding should be limited to relatively few coils which are positioned so that large quadrature fluxes are not generated. To minimize the errors due to quadrature fluxes some resolver rotors as well as stators are designed with two windings spaced effectively 90 degrees apart. By short circuiting one of these windings the errors due to quadrature fluxes can be made negligible.

The type of winding which offers the maximum number of coils for harmonic suppression and which generates no quadrature fluxes is the concentric coil-type winding which is shown in Fig. 3. The harmonic reduction factors for this winding are

$$a_n = \frac{k}{n} \sum_{j=1}^{q} N_j \sin \frac{n\beta_j}{2}$$

The distribution angle, $n\alpha_j =$ zero If the maximum number of coils is used, the pitch angle β_j is fixed and only the turns per coil are variable. For this condition the solution to equation 8 is

$$N_j = N_{\max} \sin \frac{\beta_j}{2}$$

That is, when the maximum number of concentric coils is used, a sinusoidal turns distribution produces maximum suppression of harmonics. This has been proved graphically by many designers of motors and alternators. A recent paper

Heartz, Saunders—Design of High-Precision Synchros and Resolvers
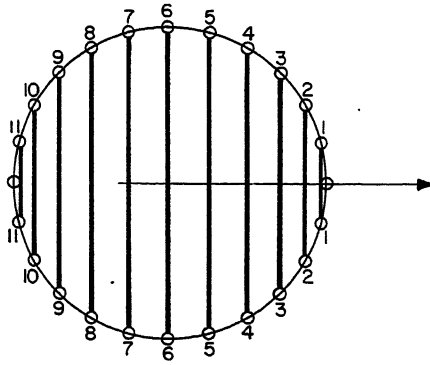
Fig. 4. Concentric coil winding

by Berman, et al.,[5] presented a general mathematical proof.

For any core which has an odd number of slots and a maximum number of coils whose turns per coil are sinusoidally distributed, all odd harmonics up to the number of slots will be zero. As the basic magnetic structure for an odd slot core cannot be bisymmetrical, even harmonics cannot be effectively cancelled. However, as the sinusoidal turns distribution presents a closely balanced winding, even harmonics will be reduced to a minimum. The equation for the harmonics which equals the number of slots will be similar to the fundamental equation, i.e.

$$\bar{a}_S = -\frac{1}{S}\,\bar{a}_1$$

where S = the number of slots. Further, the equations will be cyclical about the harmonic which equals the slot number, i.e.

$$\bar{a}_{S+2} = \bar{a}_3 = 0$$

$$\bar{a}_{S+4} = \bar{a}_5 = 0, \text{ etc.}$$

An even slot core can have a bisymmetrical winding which will remove all even harmonics. However, as one half of the winding must be identical to the other half, the effective number of coils whose turns can be varied is one-half the total number of coils. For example, a 24-slot core can have a maximum of 11 or 12 concentric coils but only six of these coils can be designed for suppression of odd harmonics. Thus turns of the 12-coil concentric winding can be designed to remove all even harmonics plus the third, fifth, seventh, ninth, and the eleventh harmonics. This winding will be cyclical about the thirteenth harmonic.

The end coils of a concentric coil winding contribute very little to the fundamental flux, whereas they do make the winding more difficult to manufacture.

If the end coils, coils 1 and 11 of Fig. 4, are omitted, then the output space harmonics will be equal to the harmonics generated by the omitted coils. The sinusoidal turns distribution is no longer effective in suppressing harmonics. The optimum turns distribution can be determined only by a complete solution of the simultaneous equations. Further, there is no guarantee that a solution which involves all positive values of turns exists. A solution, where all turns are positive, will usually exist if the third and ninth harmonic equations are dropped when the two end coils are dropped. However, in resolver applications it is essential that the third harmonic receive maximum attenuation.

An example of a winding which omits the end coils but still removes third harmonic is shown in Fig. 5. The winding equations for this winding are

$$0.500N_2 + 0.707N_3 + 0.866N_4 + 0.966N_5 = a_1$$
$$\text{(fundamental)}$$

$$1.000N_2 + 0.707N_3 + 0 - 0.707N_5 = 0$$
$$\text{(3rd harmonic)}$$

$$0.500N_2 - 0.707N_3 - 0.866N_4 + 0.259N_5 = 0$$
$$\text{(5th harmonic)}$$

$$-0.500N_2 - 0.707N_3 + 0.866N_4 + 0.259N_5 = 0$$
$$\text{(7th harmonic)}$$

The solution of these equations is: $N_2 = 0.55$, $N_3 = 0.450$, $N_4 = 0.317$, $N_5 = 1.230$. Note that to have zero third harmonic in this case, it was necessary to omit the center coil along with the two end coils.

CORE MATERIAL

In resolvers or synchros, as it is necessary that the space wave form of the air-gap flux density be sinusoidal, the flux density in various parts of the rotor and stator is not uniform. The average effective a-c permeability of a core will decrease as the flux density approaches zero or saturation. It is also a function of the core losses. When the average a-c permeability varies as a function of the magnitude of the flux density, the space wave form of the air-gap flux density is distorted. Time harmonics will also be generated in the output voltage. Another factor which causes distortion of the air-gap density is that associated with core strains. The permeability of most core materials is extremely sensitive to strains on the core. Nonuniform strains on a core will cause a variation of the a-c effective permeability which will also distort the air-gap flux density space wave form.

The magnitude of the errors which are generated by variations of the core permeability will be proportional to the mmf
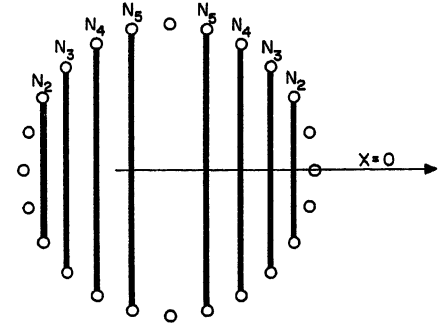
Fig. 5. Special winding design to remove third harmonic

drop which appears in the core material. If it can be assumed that the total mmf generated appears across the air gap, then variations of the core permeability will not introduce space harmonic errors. In pancake-type resolvers, the mean length of flux path in the core material is large, and the stack length is small. This requires a very small air-gap length if the phase shift between the supply voltage and the output voltage is to be less than the customary 9 degrees. These conditions lead to a large mmf drop in the core material. In one typical unit the mmf drop in a silicon iron stator was calculated to be greater than 5%.

The mmf drop in the core material can be reduced by increasing the air-gap length, by increasing the length of the rotor and stator stack or by increasing the permeability of the core. Considering these factors, the optimum core material will be 78% nickel-iron. This material has a high initial permeability (approximately 20,000 gausses per oersted) and a minimum sensitivity to strains. However, this material is soft and difficult to machine accurately. Thus, this material is best for pancake-type resolvers where the length of the stack must be small compared to the length of the flux path in the material. If the stack length can be made large compared to the length of flux path in the material then an iron-silicon alloy may be the better material. Although this alloy has a low initial permeability it is easier to machine to a tight tolerance.

MANUFACTURING TOLERANCES

The effects of manufacturing errors such as noncylindrical rotor, off-center shaft, etc., can be determined for a specific design by the method given in the latter part of the first section of this paper. However, considering the complexity of equation 6, a general treatise of manufacturing errors is beyond the

scope of this paper. Suffice it to say that because of the limits of mechanical tolerances all harmonics will be present in the air-gap flux density wave form in spite of the fact that the rotor winding is specifically designed to remove some harmonics. Practical experiences, and theoretical studies made with the aid of a digital computer have indicated that, unless both the rotor and the stator windings are carefully designed to give maximum suppression to all harmonics, it is impractical to maintain tight enough mechanical tolerances to produce a resolver unit which has less than 0.1% error.

Summary

In the design of high-precision resolvers, mechanical tolerances are limiting

factors on the accuracy. It is therefore essential that both the rotor and the stator winding have a maximum insensitivity to harmonics generated by improperly aligned rotors, nonconcentric stators etc. The concentric coil-type winding with a variable turns per coil will provide maximum harmonic suppression. A 78% nickel-iron core material offers the best magnetic characteristics for core material but this material is very difficult to machine accurately. In applications where the stator and rotor stack length can be made large as compared to the length of flux path in the core material, silicon iron alloys can be used to advantage because of their comparatively easy machineability. Although this paper is primarily devoted to design factors for high-precision resolvers, the methods of analysis and many of the results will be

valuable in the design of other types of high precision rotating devices, such as synchros, induction potentiometers, velocity generators, etc.

References

1. SELSYN INSTRUMENTS FOR POSITION SYSTEMS, T. M. Linville, J. S. Woodward. *AIEE Transactions*, vol. 53, June 1934, pp. 953–60.

2. ELECTRICAL ACCURACY OF SELSYN GENERATOR-CONTROL TRANSFORMER SYSTEM, Harold Chestnut. *Ibid.* (*Electrical Engineering*), vol. 65, Aug.-Sept. 1946, pp. 570–76.

3. STATIC ACCURACY PERFORMANCE OF THE SELSYN GENERATOR CONTROL-TRANSFORMER SYSTEM, G. Kronacher. *Ibid.*, vol. 69, pt. II, 1950, pp. 645–54.

4. HARMONICS DUE TO SLOTS IN ELECTRIC MACHINES, Robert A. Heartz, Robert M. Saunders. *Ibid.*, vol. 73, pt. III-B, Aug. 1954, pp. 946–49.

5. SPACE HARMONICS GENERATED BY CONCENTRIC WINDINGS, August Berman, Ghaffar Farmanfarma, Roland Schinzinger. *Ibid.*, Dec. 1954, pp. 1416–21.

Discussion

V. E. Hagen (Eclipse-Pioneer Division, Bendix Aviation Corporation, Teterboro, N. J.): The paper is an excellent and informative one on the effects of windings as one of the main sources of error in synchros and the means for minimizing it. However, for the past 2 decades Eclipse-Pioneer Division of Bendix Aviation Corporation has been investigating means and methods for improving the performance and fabrication of precision synchros and resolvers. Some of these important improvements are disclosed and claimed in several U. S. patents assigned to Bendix Aviation Corporation.[1-3]

At the end of the paper it is stated that "The concentric coil-type winding with a variable turns per coil will provide maximum harmonic suppression." I believe that this type of winding, among others, is disclosed and claimed in one or more of the patents mentioned. See for example, Figs. 4, 5, and 6 of references 1 and 2, and Fig. 3 of reference 3.

The statement following equation 8, "The number of harmonics which can be suppressed by a proper choice of $\bar{a}_{jn}$ is one less than the number of the coils of winding,"

is not necessarily correct any longer because of new design concepts being used by our organization.

REFERENCES

1. J. P. Glass, Jr. *U. S. Patent no. 2,488,771*, Nov. 22, 1949.

2. J. P. Glass, Jr. *U. S. Patent no. 2,535,914*, Dec. 26, 1950.

3. P. F. Bechberger et al. *U. S. Patent no. 2,550,663*, May 1, 1951.

R. A. Heartz and R. M. Saunders: The authors are well aware of the Bendix patents and the valuable work done in this field as disclosed in the Glass patent and others. However, we were concerned primarily with the scientific aspects of the problem which are not often treated in a logical fashion in letters patent. Nevertheless, the authors are grateful to Mr. Hagen for presenting this patent bibliography; it does add to the paper.

Over the years the optimum design of windings has been a subject of considerable attention in both the scientific and patent literature. The earliest treatment the authors can find is contained in a patent by Newcomb in 1899[1] illustrating a graphical method for determining the optimum wind-

ing distribution in alternating current motors. Subsequent papers by Lyon[2] (1918) and Appleman[3] (1937) are the most significant treatises in the extensive scientific literature on the subject and show that designers were well aware of the necessary conditions for obtaining minimum harmonics in electric machinery.

The statement that the number of harmonics which can be suppressed by a proper choice of $\bar{a}_{jn}$ is one less than the number of coils of the winding refers specifically to equation 8. The number of independent equations is determined by the number of independent coils. However, in general, harmonic suppression will be cyclical and the winding will suppress higher harmonics as well as those represented in equation 8, as shown by reference 5 of the paper.

REFERENCES

1. A-C MOTOR, E. C. Newcomb. *U. S. Patent no. 618,578*, Jan. 31, 1899.

2. APPLICATION OF HARMONIC ANALYSIS TO THE THEORY OF SYNCHRONOUS MACHINES, Waldo V. Lyon. *AIEE Transactions*, vol. 37, pt. II, 1918, pp. 1477–1517.

3. THE CAUSE AND ELIMINATION OF NOISE IN SMALL MOTORS, W. R. Appleman. *AIEE Transactions* (*Electrical Engineering*), vol. 56, Nov. 1937, pp. 1359–67.

Correlation Between Root-Locus and Transient Response of Sampled-Data Control Systems

ELIAHU I. JURY
ASSOCIATE MEMBER AIEE

Synopsis: The introduction of the root-locus method in the z-plane simplifies the investigation of sampled-data control systems, by demonstrating the correlation between transient response and the location of the roots. It is shown that by establishing the correlation between frequency locus and the root locus, the correlation between frequency and transient response of sampled-data systems can be demonstrated.

Synthesis of sampled-data control systems can be performed in the time domain through the introduction of phase-angle loci as demonstrated in examples. Finally, approximate equations are given so that transient overshoot and peak time may be evaluated without actually working out the transient solution.

THE performance of sampled-data control systems is evaluated on the transient-response basis when a unit step is applied, while synthesis of such systems can best be performed on the frequency-response basis.[1-4] A certain correlation exists between the maxima of the frequency response and the maxima of the transient response for second-order systems,[4] which make possible the synthesis of such systems in the frequency domain. However, it is difficult to formulate the correlation generally and quantitatively with easily understood results, especially for higher order sampled-data control systems. The introduction of the root-locus method in the z-plane makes it possible to demonstrate easily the correlation between transient-response and root location. Furthermore, by establishing the correlation between the frequency locus and the root locus, the correlation be-

tween frequency and transient response can be demonstrated. Through the introduction of phase-angle loci, the synthesis of sampled-data control systems in the time domain is possible, as indicated in examples.

The purposes of this paper are: 1. to correlate the frequency locus of the transfer function and the root-locus in the z-domain, 2. to demonstrate that synthesis of sampled-data control systems can be performed by phase-angle loci and 3. by the approximate equations given, to evaluate, without actually working out the transient solution, the transient overshoot and peak time.

Correlation Between Frequency Locus and Root Locus

The frequency locus[4] $G^*(z)$ of any sampled-data system is the locus of all points of $G^*(z)$ in the $G^*(z)$-plane as z traverses the unit circle. This locus is the conformal transformation of the whole unit circle of the z-plane on the $G^*(z)$-plane. Fig. 1 indicates a typical sampled-data control system whose frequency locus for various values of gain K is shown in Fig. 2. The z-transform input-output of such a system, assuming $N^*(z)$ is unity, is given[3] as follows

$$\theta_c*(z)=\theta_r*(z)\frac{G^*(z)}{1+G^*(z)} \tag{1}$$

where $G^*(z)$ is the z-transform of the open-loop transfer function and $\theta_c*(z)$, and $\theta_r*(z)$ are the z-transforms of the output and input respectively.

The root locus is the locus of all values of z when substituted in $G^*(z)$ gives the function $G^*(z)$, a phase angle of $(2m-1)\pi$ where m is integer number. It is also the conformal transformation of the negative real axis of the $G^*(z)$-plane on the z-plane. Fig. 3 indicates a typical root locus of $1+G^*(z)$. In brief, the frequency loci are certain conformal transformations from the z-plane to the $G^*(z)$-plane or from $G^*(z)$-plane to the z-plane respectively by the complex function $G^*(z)$.

This transformation is regarded as the correlating relationship between the frequency locus and the root locus.

The root locus is actually a graphical method for obtaining the poles of the overall transfer functions. Thus, if a sampled-data control system is characterized by equation 1, then the poles of the overall transfer function will evidently be all those values of z which make $G^*(z) = -1$. In seeking to find values of z which make the transfer function $G^*(z)$ equal to -1, the value -1 is considered as a vector whose angle is $(2m-1)\pi$ and magnitude is unity. First, the problem may be considered of finding the locus for which the angle condition alone is satisfied, i.e., plot the locus of z for the open-loop transfer function $G^*(z)$ so that the addition of all angles of the various terms of $G^*(z)$ add up to $(2m-1)\pi$. This locus can be easily plotted using a spirule[5] or other procedures indicated later. After the locus has thus been determined, one considers the second condition, i.e., that the magnitude of $G^*(z)$ be unity at a particular value of z which corresponds to $(2m-1)\pi$. This procedure determines the maximum allowable value or range of values of the gain K, which makes the sampled-data system stable or causes it to have a certain overshoot and peak-time response.

Since the frequency locus is the conformal transformation of the unit circle in the z-plane, the frequency locus itself cannot give the value of the poles. Any attempt to find the poles by means of the correspondence between the z- and $G^*(z)$-planes requires some transformation along a certain curve inside the unit circle where the desired poles are located. Thus, the frequency locus may be generalized by putting

$$z=A_he^{j\phi} \tag{2}$$

into the transfer function $G^*(z)$. (For a stable system the particular values of z to be considered should lie inside the unit circle.) Using A_h as a parameter, a family of constant A_h locus is obtained. The usual frequency locus is the constant A_h locus, where $A_h =$ unity. If the radial lines from the origin of the z-plane are conformally transformed on the $G^*(z)$-plane, a family of constant ϕ locus is obtained. By means of these two families of loci, a curvilinear square in $G^*(z)$-plane can be found which encloses the point -1, 0, and by taking successively close approximation in the values of z the poles of equation 1, can be found to any desired accuracy. Typical families of constant A_h and constant ϕ and their conformal transformation in the $G^*(z)$-plane are

Fig. 1. Typical sampled-data control system

shown in Figs. 4 and 5. This generalized frequency locus is included in this paper only as an indication of the possible methods of obtaining the poles of equation 1.

The root locus may be generalized by the relation

$$\text{Ang } G^*(z) = \psi \tag{3}$$

where ψ may be any angle varying from zero to $\pm \pi$. For various values of ψ, a family of constant ψ of constant phase-angle loci may be constructed which are analogous to those used in continuous systems.[6]

Typical constant phase-angle loci will be derived as follows.
For case 1

$$G^*(z) = 1/z \tag{4}$$

Since z is complex, which can be written as $x + jy$; thus $\psi = -\tan^{-1} y/x$, and the family of constant ψ loci has the following:

$$y = (-\tan \psi)x \tag{5}$$

Equation 5 indicates that constant ψ loci are straight lines as shown in Fig. 6.
For case 2

$$G^*(z) = 1/(z-1) \tag{6}$$

In this case $\psi = -\tan^{-1} y/x - 1$, and the equation of constant phase loci is

$$y = (-\tan \psi)x + \tan \psi \tag{7}$$

Equation 7 is a family of straight lines as shown in Fig. 7.
For case 3

$$G^*(z) = z - b/z - a \tag{8}$$

The phase angle ψ in this case can be easily shown as

$$\psi = \tan^{-1} y/x - b - \tan^{-1} y/x - a \tag{9}$$

By putting $N = \tan \psi$ in equation 9, the constant ψ loci equation after certain algebraic manipulation can be written as follows

$$\left(x - \frac{a+b}{2}\right)^2 + \left(y - \frac{b-a}{2N}\right)^2 = \left(\frac{a-b}{2}\right)^2 + \left(\frac{b-a}{2N}\right)^2 \tag{10}$$

Equation 10 is a family of circles as shown in Fig. 8 for a and b real numbers. It is noted that the root locus is the constant ψ locus, for which ψ is equal to $(2m-1)\pi$. Since the z-transform of the transfer function, $G^*(z)$ is generally regarded to consist of combinations of the mentioned three cases, thus the three

basis forms of $G^*(z)$ discussed in this paper will suffice to cover most forms of $G^*(z)$. To obtain the roots of $1 + G^*(z)$ in equation 1, one can apply the magnitude condition, i.e., $|G^*(z)| = 1$, which yields the actual roots.

The frequency locus $G^*(z)$ can also be obtained from the root locus once the roots have been determined, in the same manner as obtaining the frequency locus from the root locus in the continuous systems.

From the correspondence between z and the $G^*(z)$-plane, any information of the system encompassed on the $G^*(z)$-plane may be obtained from the z-plane by means of certain transformation. To find the actual correlation between frequency and transient responses, advantage can be taken of the fact that the loci on the $G^*(z)$-plane give information on the frequency response, and the loci on the z-plane on the transient response.

Synthesis of Sampled-Data Control Systems in Time Domain

Shaping of $G^*(z)$ to yield favorable over-all frequency response can be accomplished by insertion of linear networks either in the forward or the feedback path. The M (or the frequency overshoot) criterion[4] will assure a satisfactory transient response if the attenuation diagram of the closed-loop frequency response is examined. Frequency shaping of $G^*(z)$ by inserting linear networks is inherently complicated because the magnitude and phase angle of the open-loop frequency response cannot be obtained by the product of magnitudes and summation of phase angles of each cascaded component of the system in $G^*(z)$. However, when digital computers[7] or pulsed networks are inserted in the open loop as shown in
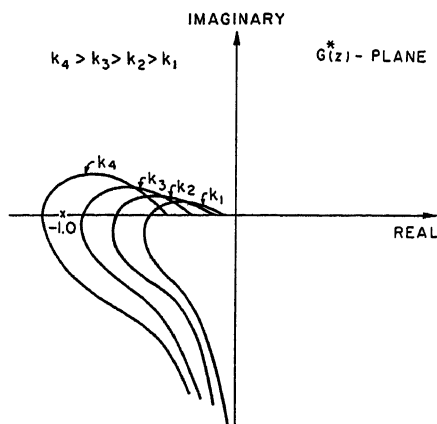
Fig. 2 (left). Typical frequency focus of $G^*(z)$
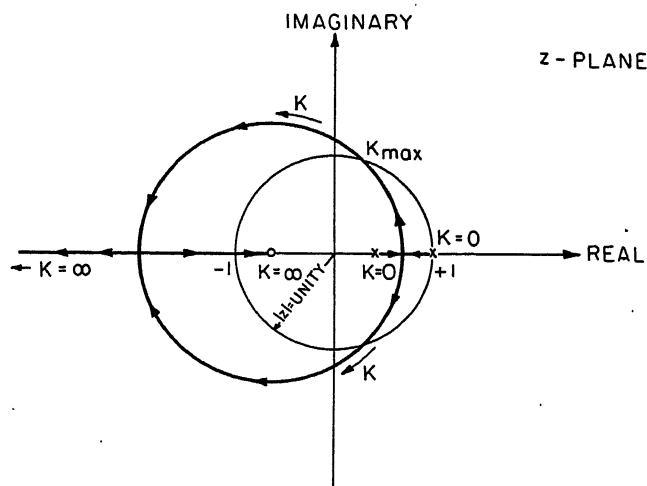
Fig. 3 (right). Typical root-locus of $1 + G^*(z)$
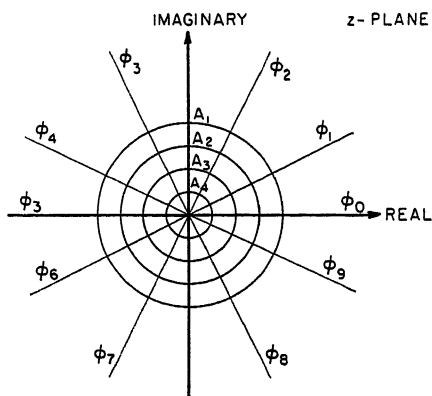
Fig. 4 (left). Family of constant ϕ locus and constant A_h locus in the z-plane
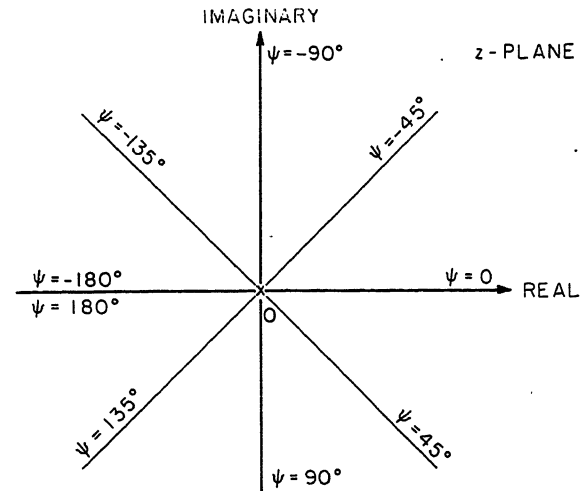
Fig. 6 (right). Phase-angle loci of $1/z$

Fig. 1, then by multiplication of the magnitudes and the addition of the phase angles of the frequency loci of the components, the open-loop frequency response can be obtained and shaping of $G^*(z)$ is readily performed in very much the same way as for continuous systems.

Sampled-data control systems can best be synthesized in the time domain by the use of phase-angle loci, provided pulsed networks or discrete sample stabilizers are employed. As will be shown later, that accurate relation between the chosen roots from the root locus and the transient response are readily obtained, so that transient response may be evaluated quickly without actually working the transient solution in such cases. The root locus can be constructed by superimposing the phase-angle loci of the cascaded components of $G^*(z)$ on the z-plane and identifying the points which give a sum of $(2m-1)\pi$. The choice of the proper poles of equation 1 from the root locus will yield the desired transient response.

The shaping of the root locus to yield the desired roots can be accomplished by adjusting the constants of the stabilizing components in the loop which might be digital computer or pulsed networks. By so doing, a qualitative picture may be obtained of how the original root locus tends to change. In view of the fact that all sampled-data control systems are only conditionally stable,[4] the problem of shaping the root locus is more difficult than in the continuous case, hence trial-and-error procedures are more prevalent in the synthesis of such systems than in the continuous ones. The compensating effect of pulsed networks is limited by the actual practical considerations of the range of constants. However, digital computers can be utilized more freely for they are subjected to fewer restrictions.

ILLUSTRATIVE EXAMPLE

If a sampled-data system such as is

shown in Fig. 1 is assumed (without the stabilizing network) and with

$$G(s) = \frac{K}{s(s+a)} \qquad (11)$$

it can be easily shown that the corresponding $G^*(z)$ of this second-order system with zero hold circuit is

$$G^*(z) = \frac{KT}{a(z-1)} - \frac{K(1-\epsilon^{-aT})}{a^2(z-\epsilon^{-aT})} \qquad (12)$$

The root-locus equation can be obtained by letting ψ, the phase angle of $G^*(z) = \pi$, from which, see Appendix, is derived the following

$$y^2+(x+b)^2 = b^2+b+\epsilon^{-aT}(b+1) \qquad (13)$$

$$b = \frac{1-\epsilon^{-aT}-aT\epsilon^{-aT}}{aT-1+\epsilon^{-aT}} \qquad (14)$$

Equation 13 is a circle in the z-plane which indicates part of the root locus. These circles have been plotted for various values of T as shown in Fig. 9. It is noticed that the roots for maximum allowable gain are complex when T is small, and real when "T" becomes larger, and the transition between those types of poles is when $T=3.7$. Therefore, the system can oscillate at two frequencies: 1. when $\omega = \pi/T$, and 2. when $\omega < \pi/T$, as has been deduced in previous work.[2,4]

As an illustration, two root loci are plotted in Figs. 10 and 11, when $T=2$ and 5 respectively and $a =$ unity in equation 13.

If the values $T=1$, $K=1$, $a=1$ are inserted in equation 12, then

$$G^*(z) = \frac{0.368}{z-1} \frac{z+0.72}{z-0.368} \qquad (15)$$

The constant-phase loci of equation 15 are obtained by superimposing the various phase-angle loci of the two components of $G^*(z)$, i.e., the straight lines and circles.
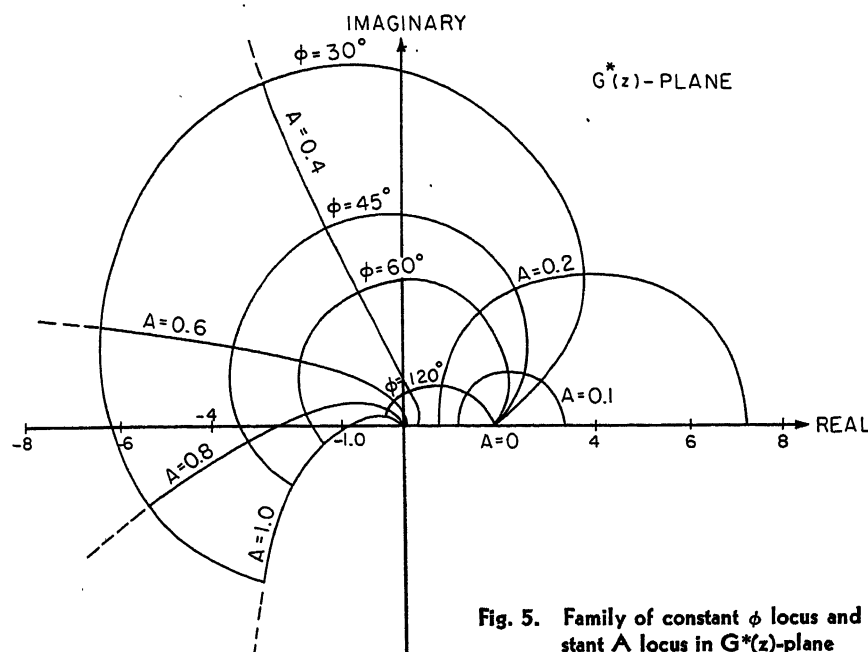
Fig. 5. Family of constant ϕ locus and constant A locus in G*(z)-plane
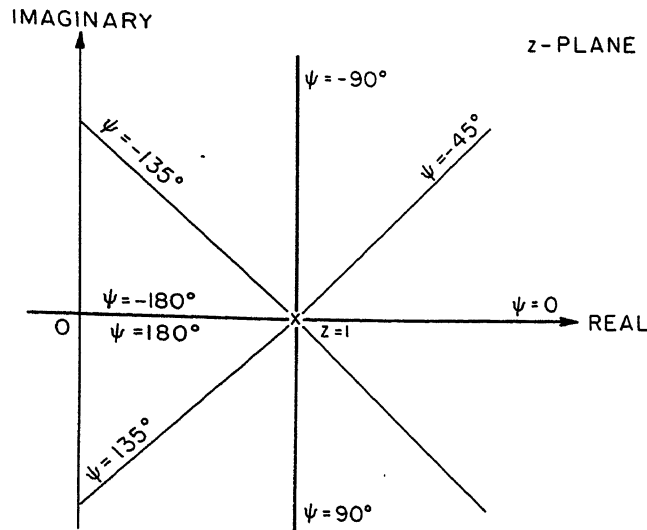
Fig. 7 (left).
Phase-angle loci
of $1/z$-1
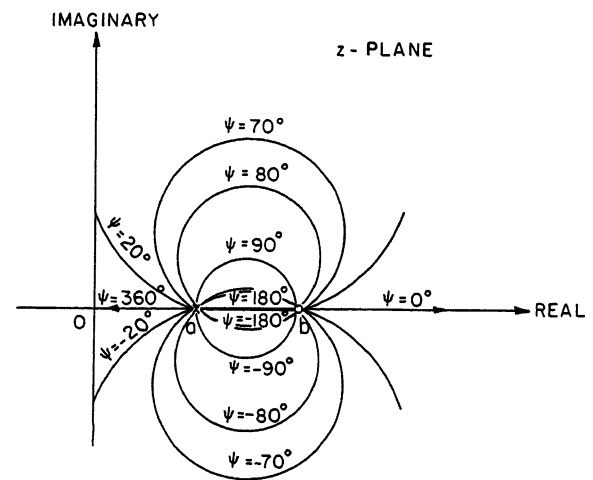
Fig. 8 (right).
Phase-angle loci
of z-b/z-a

The root locus is obtained by identifying the points which give a sum of ± 180 degrees. The various constant-phase loci and the root locus of equation 15 is constructed in Fig. 12.

Part of the root-locus equation can be obtained from equation 15 as

$$y^2+(x+0.72)^2=(1.368)^2 \qquad (16)$$

Equation 16 denotes a circle of radius (1.368) and the center at $(-0.72, 0)$ which agrees with the graphical construction in Fig. 12. To find K from the root locus, the equation

$$K=\left|\frac{(z-1)(z-0.368)}{z+0.72}\right|\frac{1}{0.368} \qquad (17)$$

should be satisfied,
when $K=$unity the roots are

$$z_{1,2}=0.5\pm j\sqrt{0.38} \qquad (18)$$

The maximum allowable gain K_m is obtained at the intersection of the root locus with the unit circle, at which point the roots are

$$z_{1,2}=0.24\pm j0.97 \qquad (19)$$

K_m is equal to 2.43, which indicates the limit of stability of this particular system.

Suppose the system described in equation 11 is to be compensated with the pulsed-lead network, or digital computer, as shown in Fig. 1; then the transfer function $N^*(z)$ of the compensating system is given as

$$N^*(z)=A\frac{z-B}{z-C} \qquad (20)$$

The choice of the digital compensation in the form given in equation 20 will further stabilize the system if $B>C$, as may be noticed from the constant phase loci of this function and the shaped root locus in Fig. 13.

Assume

$$A=0.5,\ B=0.736,\ C=0.368 \qquad (21)$$

Equation 20 becomes

$$N^*(z)=0.5\left(\frac{z-0.736}{z-0.368}\right) \qquad (22)$$

The compensated open-loop z-transform is then

$$N^*(z)G^*(z)=0.5\left(\frac{z-0.736}{z-0.368}\right)\left(\frac{z+0.72}{z-0.368}\right)\times\frac{0.368K}{z-1} \qquad (23)$$

The root locus of $1+N^*(z)\ G^*(z)$ is as shown in Fig. 13, and the value of the gain can be found as follows

$$K=\frac{2}{0.368}\left|\frac{(z-1)(z-0.368)}{(z+0.72)}\left(\frac{z-0.368}{z-0.736}\right)\right| \qquad (24)$$

For the same distance from the origin as for the roots of the uncompensated locus when $(K=1)$, the roots

$$z_{1,2}=0.13\pm j\ 0.79,\ z_3=+0.766 \qquad (25)$$

and $K=3.81$.

The limiting value of gain K_m in this case can be found when the shaped root locus crosses the unit circle, which yields the following roots

$$z_{1,2}=-0.1\pm j\ 0.995,\ z_3=+0.744 \qquad (26)$$

The maximum gain before instability is reached is

$$K_m=6.48 \qquad (27)$$
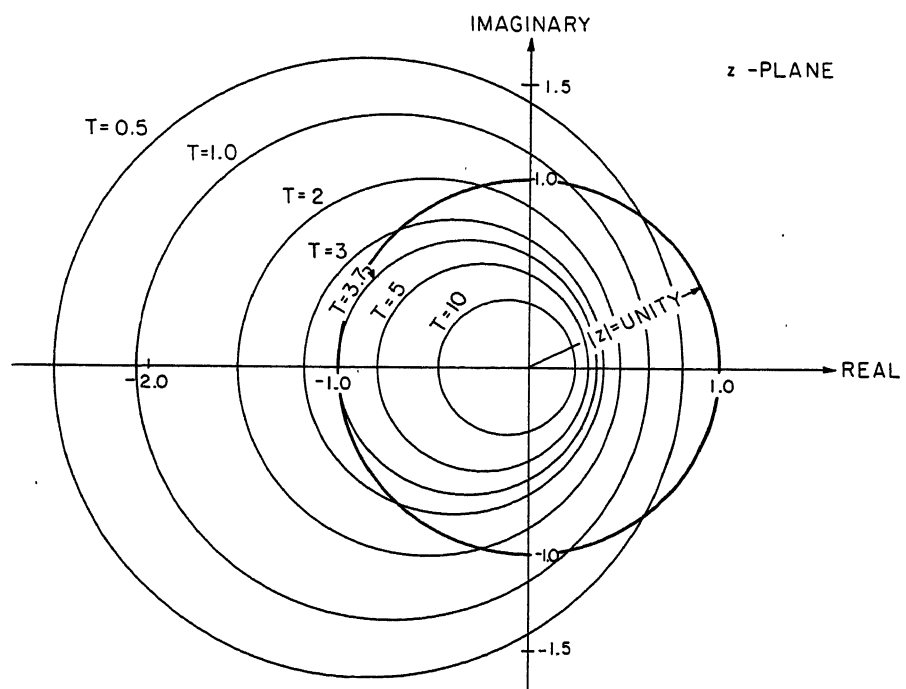
Fig. 9. Family of root-locus plots of system in example for various values of T
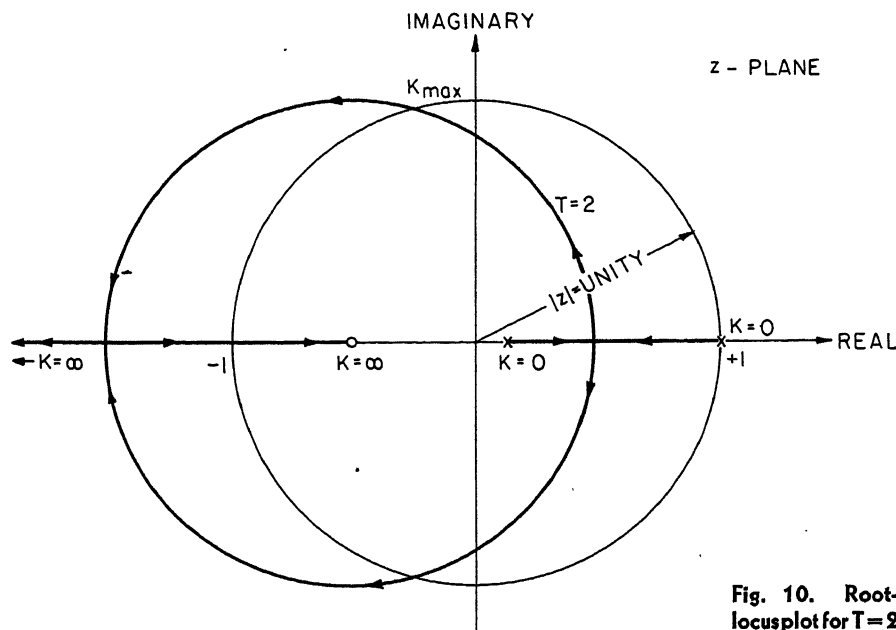
Fig. 10. Root-locus plot for T=2

and that all other poles are concentrated near the origin of the unit circle. Then all the modes of equation 33 may be neglected except the constant term and those due to the pair of predominate poles $(\alpha_0 \pm j\beta_0)$, thus

$$C_{(nT)} = K\frac{A(1)}{B(1)} + 2\left|\frac{KA(z_0)}{(z_0-1)B'_{(z_0)}}\right||z_0|^n \times$$
$$\cos\;[n\theta_0 + \text{Ang}\;A(z_0) - \text{Ang}(z_0-1) -$$
$$\text{Ang}\;B'_{(z_0)}]\quad (35)$$

where

$$z_0 = |\alpha_0 + j\beta_0| \quad (36)$$

$$B(z_0)' = \frac{dB}{dz}\Big|_{z=z_0} \quad (37)$$

Suppose the response of such a system is as shown in Fig. 15, where M_p and n_p are the important quantities that characterize this response. To obtain n_p, apply the following condition and solve for n_p which ought to be integer.

$$C(n+1) - C(n) = 0 \quad (38)$$

The solution of equation 38 generally yields the value of n which is noninteger; however, the actual value of n_p is the upper integer value of the actual n obtained. In the following derivations, it is assumed that the solution of equation 38 yields an integer value of n.

Substituting equation 35 in 38, it can be shown that

$$n_p = \frac{1}{\theta_0}\left[\frac{\pi}{2} - \text{Ang}\;A(z_0) + \text{Ang}\;B'_{(z_0)}\right]\quad (39)$$

The first overshoot is obtained while substituting equation 39 in 35, thus

It is noticed that the compensated system is more stable than the original one, but the values of gain are limited for stability even in the compensated case. The results are compared in Table I.

Approximate Equations for Transient-Overshoot and Peak-Time

For a sampled-data control system as shown in Fig. 14, assume that the input-output z-transform relation can be represented as follows

$$\frac{C^*(z)}{R^*(z)} = \frac{G^*(z)}{1+HG^*(z)} = K\frac{A(z)}{B(z)} \quad (28)$$

where

$$A(z) = \prod_{i=1}^{m}(z-p_i)$$

$$B(z) = \prod_{k=1}^{m}(z-q_k) \quad (29)$$

It is assumed that no multiple poles of $C^*(z)/R^*(z)$ exist and that the order of $B(z)$ in z is higher than $A(z)$, as encountered in practical systems. For step-input

$$R^*(z) = \frac{z}{z-1} \quad (30)$$

Substituting equation 30 in equation 28

$$C^*(z) = K\frac{z}{z-1}\frac{A(z)}{B(z)} \quad (31)$$

The inverse z-transform of $C^*(z)$ is

$$C(nT) = \frac{1}{2\pi j}\int_{\Gamma}K\frac{A(z)}{(z-1)B(z)}z^n dz \quad (32)$$

where Γ is a contour of integration in the z-plane that encloses all the singularities

of the integrand in equation 32. Performing the integration, it can be shown that equation 32 is

$$C_{(nT)} = K\frac{A(1)}{B(1)} + \sum_{q_k = \text{real}}\frac{KA(q_k)}{(q_k-1)B'_{(q_k)}}\;(q_k)^n +$$

$$\sum_{q_k = \alpha_k + j\beta_k}2\left|\frac{KA(q_k)}{(q_k-1)B(q_k)}\right|[\alpha_k^2 + \beta_k^2]^{n/2}\times$$
$$\cos\;[n\theta_k + \text{Ang}\;A(q_k) - \text{Ang}\;(q_k - 1) - \text{Ang}\;B'_{(q_k)}]\quad (33)$$

where

$$\tan\theta_k = \beta_k/\alpha_k \text{ and } B'_{(q_k)} = \frac{dB}{dz}\Big|_{z=q_k} \quad (34)$$

Assume that a pair of predominate poles exists $(\alpha_0 \pm j\beta_0)$ near the unit circle

Fig. 11. Root-locus plot for T=5
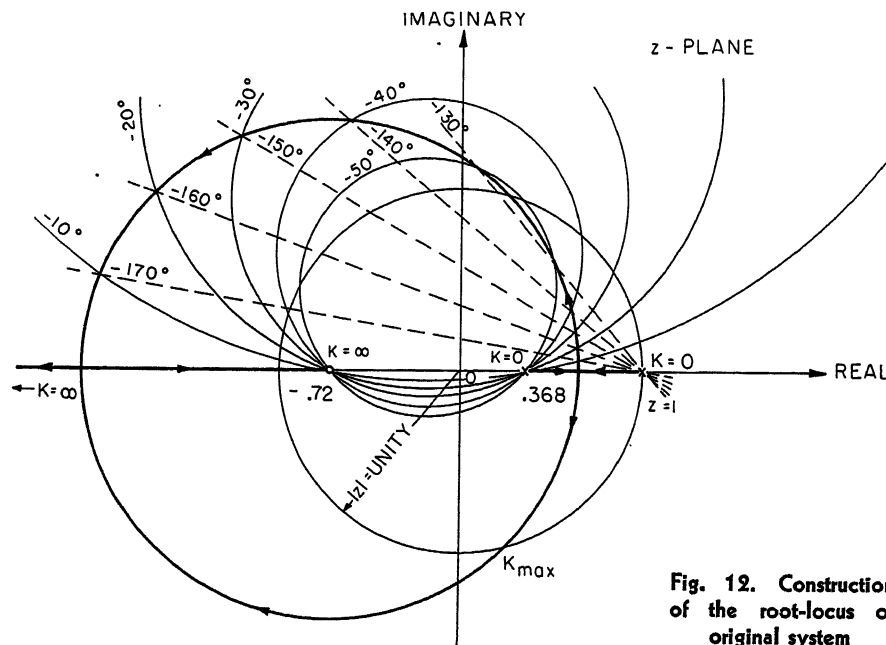
Fig. 12. Construction of the root-locus of original system

Term	Symbol	Original	Compensated
Complex roots..	$z_{1,2}$	..$0.5\pm j\,0.38$..	$0.13+j0.79$
Real roots......	z_3	....	0.766
Distance of complex roots from origin of unit circle....	$\sqrt{z_1 z_2}$	0.795	0.795
Gain..........	K ..	1 ..	3.81
Maximum gain..	K_m ..	2.43 ..	6.4
Peak overshoot, per cent......	M_p ..	39.9 ..	40.7
Peak time, seconds.........	n_p ..	3 ..	2

$$M_{\text{peak}} = \frac{2}{[(\alpha_0-1)^2+\beta_0^2]^{1/2}}\left|\frac{KA(z_0)}{B'(z_0)}\right|\times$$
$$|z_0|^{n_p}\sin\,\text{Ang}(z_0-1)] \quad (40)$$

Equation 40 can be simplified further as follows

$$M_p = \frac{2\beta}{[(\alpha_0-1)^2+\beta_0^2]}\left|\frac{KA(z_0)}{B'(z_0)}\right||z_0|^{n_p} \quad (41)$$

The accuracy of the assumption which resulted in the two approximate relations for n_p and M_p may be explained readily. From the general solution of equation 33, a certain simple mode is considered; that is, the mode

$$\frac{KA(q_1)}{(q_1-1)B'(q_1)}(q_1)^n$$

This mode will die to a small per cent of its initial value when $n_p=3$, and if n_p is larger; then the effect of this simple pole is small even if

$$\frac{KA(q_1)}{(q_1-1)B'(q_1)}$$

is large. It can be shown that the coefficient

$$\frac{KA(q_1)}{(q_1-1)B'(q_1)}$$

is small if this root q_1 is not too close to

Fig. 13 (left). Construction of the root-locus of compensated system

Fig. 14 (below). Sampled-data control system

other roots. In case a certain mode must be accounted for, the magnitude of this mode at n_p can be evaluated from the proper term of the general equation 33; this is added to the result of equation 35 as a second approximation.

Equation 39 can be rewritten as

$$n_p\theta_0 = \frac{\pi}{2}-(\text{sum of angles from zeros to}$$
the predominate pole $\alpha_0+j\beta_0$)
$+$(sum of angles from other poles to the predominate pole
$$\alpha_0+j\beta_0) \quad (42)$$

Thus it may be concluded that zeros decrease, n_p, and additional poles increase n_p. Furthermore, n_p is inversely proportional to θ_0.

M_p, from equation 41 can be written as the product of the following two terms

$[\alpha_0^2+\beta_0^2]^{n_p/2}$ (This quantity is always less than unity, since the predominate poles are inside the unit circle.)

$$\frac{K}{(\alpha_0-1)^2+\beta_0^2}\times(\text{product of distances}$$
from zeros to the predominate pole $\alpha_0+j\beta_0$)/(product of distances from all poles to the predominate pole $\alpha_0+j\beta_0$, excluding the distance between the two predominate poles. (43)

It is evident from the first term that the smaller n_p, the larger M_p, thus the choice of n_p and M_p requires a compromise. For a zero-displacement error system K must fulfill the condition that in steady

Fig. 15 (left). The desired response

Fig. 19 (right). Effect on M_p and n_p when $a0 < ab$
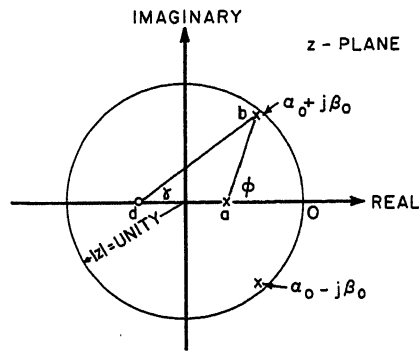
Fig. 16. Location of predominate poles in the z-plane
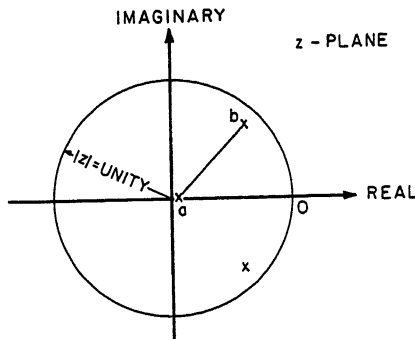
Fig. 17. Effect on M_p and n_p when a third root is near the origin of the unit circle in the z-plane
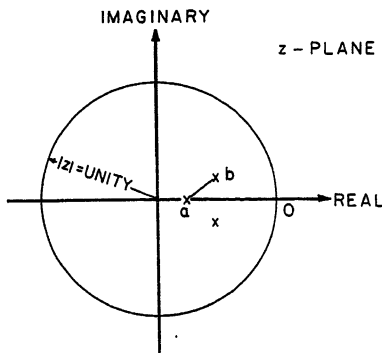
Fig. 18. Effect on M_p and n_p when $a0 > ab$

state the output equals the input, thus

$$K = \frac{\prod\limits_{k=1}^{n} (1-q_k)}{\prod\limits_{i=1}^{m} (1-p_i)} \qquad (44)$$

Substituting for K into equation 41 to obtain

$M_{\text{peak}} = $ (Product of distances from all poles to the point (1,0), excluding distances from two predominate poles to the point 1,0)/(product of distances from all poles to predominate pole $\alpha_0 + j\beta_0$, excluding the distance between predominate poles)× (product of distances from zeros to the predominate pole $\alpha_0 + j\beta_0$)/(product of distance from all zeros to the point

$$1,0)|z_0|^{n_p} \qquad (45)$$

In equation 38 n_p can be written as

$$n_p = \frac{1}{\theta_0}\left(\frac{\pi}{2} - \sum_{\text{all zeros}} \gamma + \sum_{\substack{\text{all poles except} \\ \text{predominate pair}}} \phi \right)$$

$$(46)$$

where ϕ and γ are as shown in Fig. 16. Substituting for n_p in equation 45, then

$$M_{\text{peak}} = \left[\pi \text{ (all poles except predominate pair) } \frac{a0}{ab}|z_0|^{\phi/\theta_0} \text{ (due to additional poles)} \right] \times \left[\prod \text{(all zeros)} \frac{db}{d0} |z_0|^{-\gamma/\theta_0} \text{(due to zeros)} \right] \times$$

$$[|z_0|^{\gamma/2\theta_0} \text{ (constant term)]} \qquad (47)$$

where $a0$, ab, db, $d0$ are the distances as shown in Fig. 16.

In discussing the effect of real poles on M_p, it is evident from Fig. 17, that if a pole is near the origin of the unit circle,

Fig. 19 (right). Effect on M_p and n_p when $a0 < ab$
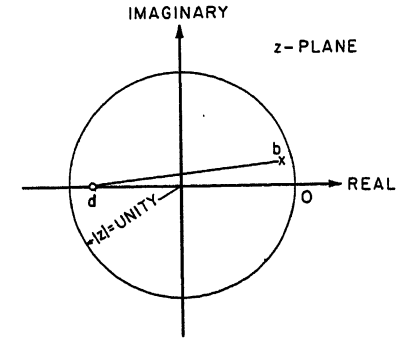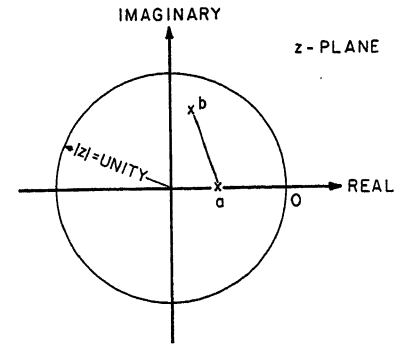
Fig. 20. Effect on M_p and n_p when $db/d0 \approx 1$ and γ is small
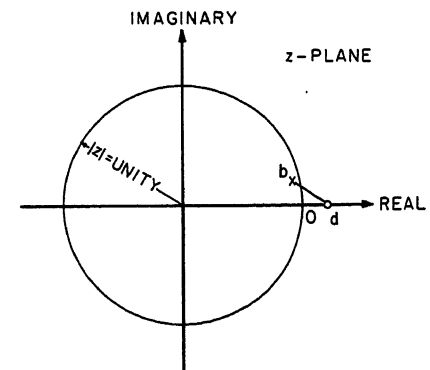
Fig. 21. Effect on M_p and n_p when a zero exists outside the unit circle

then $a0/ab$ is nearly unity and the effect of this pole configuration on both M_p and n_p is not substantial. In contrast to this, if a pole configuration is as shown in Fig. 18, then $a0/ab > 1$, and M_p will increase considerably. Thus it can be deduced that the response will be highly oscillatory if the poles are coinciding. Finally, if a pole configuration is as shown in Fig. 19, then $a0/ab < 1$, and M_p is reduced, compared to the case of Fig. 17. Similar conclusions can be deduced when the additional poles are complex inside the unit circle.

The effect of zeros on M_p and n_p can be deduced from the second bracket of equation 47, where it is noticed that for a zero configuration as shown in Fig. 20,

$db/d0 \approx 1$, and γ is small. The effects on both M_p and n_p are small. In contrast, when a zero is near the point $(1,0)$, then $db/d0 > 1$, and M_p is increased tremendously; thus a zero near the point $(1,0)$, is not desirable. However, when a zero is outside the unit circle, as shown in Fig. 21, then $db/d0 \approx 1$, and M_p will not increase considerably. In conclusion, it ought to be mentioned that real and complex poles might exist in any location inside the unit circle, while zeros of the over-all transfer function may be located anywhere inside or outside the unit circle.

Conclusions

By establishing the correlation between the frequency locus on the $G^*(z)$-plane and the root locus on the z-plane through conformal transformation, the frequency response may be interpreted in terms of the parameters of the transient response.

The introduction of phase-angle loci and the superposition method, make possible the plotting and shaping of the root locus to be accurately achieved. The reshaping can be achieved practically for any stabilizer element of the system, provided it is a pulsed network or a digital computer, the parameters of which may be varied. The roots chosen from the root locus, together with zeros, will determine the transient response. By developing approximate equations when a pair of predominate poles exist, a system may be synthesized without actually working out the transient response. The approximate relations for peak-time n_p and overshoot M_p may be used to determine the effect of the location of zeros as well as other poles on peak-time n_p and peak overshoot M_p.

Thus it is indicated that synthesis of sampled-data control systems can be performed on the root locus as well as on the frequency locus. The choice between the two lies in the convenience in plotting the root locus and in the particular results to be sought. This convenience is emphasized more in those systems using pulsed networks or digital computers than in those using networks in the continuous element, in which case the frequency locus is best applicable.

Appendix. Derivation of Root-Locus Equation of Second-Order Sampled-Data Control System

The transfer function of a second-order system with hold circuit is as follows

$$G(s) = \frac{K}{s(s+a)} \frac{(1-\epsilon^{-Ts})}{s} \tag{48}$$

The corresponding z-transform of $G(s)$ is

$$G^*(z) = \frac{KT}{a(z-1)} - \frac{K}{a^2} \frac{1-\epsilon^{-aT}}{z-\epsilon^{-aT}} \tag{49}$$

or

$$G^*(z) = \frac{K(T-1/a+\epsilon^{-aT}/a)}{a} \times \frac{z+b}{(z-1)(z-\epsilon^{-aT})} \tag{50}$$

where

$$b = \frac{1-\epsilon^{-aT} - Ta\epsilon^{-aT}}{aT-1+\epsilon^{-aT}} \tag{51}$$

Since $z = x+jy$, thus $G^*(z)$ can be written as

$$G^*(z) = \frac{K(T-1/a+\epsilon^{-aT}/a)}{a} \times \frac{(x+b)+jy}{[(x-1)+jy][x-\epsilon^{-aT}+jy]} \tag{52}$$

The phase of $G^*(z)$ in equation 52 can be easily shown to be

$$\psi = \tan^{-1} y/x+b - \tan^{-1} \times \frac{y(x-1)+(x-\epsilon^{-aT})}{(x-1)(x-\epsilon^{-aT})-y^2} \tag{53}$$

Putting $N = \tan\psi$ in equation 53, it can be shown that

$$N = \tan\psi = \frac{y/x+b - \left[\dfrac{y\{(x-1)+(x-\epsilon^{-aT})\}}{(x-1)(x-\epsilon^{-aT})-y^2}\right]}{1+(y/x+b)\left[\dfrac{y(2x-1-\epsilon^{-aT})}{(x-1)(x-\epsilon^{-aT})-y^2}\right]} \tag{54}$$

To obtain the root-locus $\psi = \pi$, or $N = 0$, thus equation 54 becomes

$$\frac{1}{x+b} - \frac{(2x-1)-\epsilon^{-aT}}{(x-1)(x-\epsilon^{-aT})-y^2} = 0 \tag{55}$$

which finally yields

$$y^2+(x+b)^2 = b^2+b+\epsilon^{-aT}(b+1) \tag{56}$$

References

1. SAMPLED-DATA CONTROL SYSTEMS STUDIED THROUGH COMPARISON OF SAMPLING WITH AMPLITUDE MODULATION, William L. Linvill. *AIEE Transactions*, vol. 70, pt. II, 1951, pp. 1779–88.

2. THEORY OF SERVO-MECHANISM (book), Hubert M. James, Nathaniel B. Nicholls, R. S. Phillips. Massachusetts Institute of Technology Radiation Laboratory Series, vol. 25, 1947, chap. 6.

3. THE ANALYSIS OF SAMPLED-DATA SYSTEMS, J. R. Ragazzini, L. A. Zadeh. *AIEE Transactions*, vol. 71, pt. II, Sept. 1952, pp. 225–234.

4. ANALYSIS AND SYNTHESIS OF SAMPLED-DATA CONTROL SYSTEMS, Eliahu I. Jury. *AIEE Transactions*, vol. 73, pt. I, Sept. 1954, pp. 332–46.

5. CONTROL SYSTEM SYNTHESIS BY ROOT LOCUS METHOD, Walter R. Evans. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 66–69.

6. SYNTHESIS OF FEEDBACK CONTROL SYSTEMS BY PHASE-ANGLE LOCI, Yaohan Chu. *AIEE Transactions*, vol. 71, pt. II, Nov. 1952, pp. 330–39.

7. ANALYSIS OF CONTROL SYSTEMS INVOLVING DIGITAL COMPUTERS, W. K. Linvill, J. M. Salzer. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 41, no. 7, July 1953.

8. SYNTHESIS OF CLOSED LOOP SYSTEMS USING CURVILINEAR SQUARES TO PREDICT ROOT LOCATION, D. W. Russell, C. H. Weaver. *AIEE Transactions*, vol. 71, pt. II, Jan. 1952, pp. 95–104.

9. THE EFFECT OF POLE AND ZERO LOCATIONS ON THE TRANSIENT RESPONSE OF SAMPLED-DATA SYSTEMS, Eliahu I. Jury. *AIEE Transactions*, vol. 74, pt. II, Mar. 1955, pp. 41-48.

———◆———

Discussion

Masahiro Mori (University of Tokyo, Chiba City, Japan): I also proposed the root-locus method of the pulse transfer function last year in Japan[1] and I agree with the views that this method is convenient for those systems using pulsed network which are important for good control.

It can be shown that the rules[2] used to plot the root-loci of continuous systems are also applied to these sampled-data systems, because both systems have transfer functions of the same rational function forms. The use of the phase-angle loci is a good method to plot the root-locus of the function. as in equation 15, which has two components. But it seems very complicated when the transfer function has more than two components, and not any more convenient than using the rules.

Approximate equations for M_p and n_p, and the effect of the location of zeros and poles on M_p and n_p expressed in this paper are valuable for synthesis and optimum adjustment. I have proved the relation between the location of roots and the shapes of initial response, as shown in Table II. At the same time I found the damping criterion on the z-plane.

Another form of equation 33 is

$$\begin{aligned} C_{(nT)} = &(\alpha_1\zeta_1{}^n + \alpha_2\zeta_2{}^n + \ldots) + \\ &[(\beta_1\eta_1{}^n \sin n\phi_1 + \gamma_1\eta_1{}^n \cos n\phi_1) + \\ &(\beta_2\eta_2{}^n \sin n\phi_2 + \\ &\qquad \gamma_2\eta_2{}^n \cos n\phi_2) + \ldots] \end{aligned} \tag{57}$$

where α_i, β_i, γ_i are constants, and ζ_i, η_i are the constants which gave values according to locations of roots. The first term may be introduced from real roots and the second from conjugate complex roots. From equation 57, the result as shown in Table II may be deduced.

Now, the output sequence which is introduced from a pair of conjugate roots is

$$C_{(nT)} = 2|q_k|^n |R_k| \cos(n\phi_k + \text{Ang } R_k) \tag{58}$$

where $|q_k|$ = absolute value of root, and R_k = residue of q_k. If n_0 is defined as

$$n_0 = \frac{2\Pi}{\phi_k} \tag{59}$$

for instance, at the case of 25-per-cent (%) damping

$$\frac{C(n_0T)}{C(0)} = \frac{1}{4} \tag{60}$$

hence

$$|q_k| = \frac{1}{4} \tag{61}$$

Table II. Correspondence Between z-Plane and s-Plane

Stability criterion......inside of unit circle............	left half-plane
One uniform variation....one positive real root............	one negative real root
One damping..one pair of conjugate complex roots or one negative real root............	one pair of conjugate complex roots
Critical damping.....intersection with positive real axis............	intersection with negative real axis
25% damping..heart-shaped curve in Fig. 22(A)..........	straight lines in Fig. 22(B)

spiral for $\phi_k > \Pi$ is unnecessary. Then the heart-shaped curve in Fig. 22(A) is the condition for 25% damping. If the root is located on this curve, the damping is 25%. When the root is located inside this curve, the damping is stronger than 25%, and when outside, the damping is weaker. The optimum condition may be determined by finding the value of K at the intersection of this curve with the root locus.

It is also evident that the criteria in the z-plane are the conformal representations of those in the s-plane, with such a condition as $z = \epsilon^{sT}$; see column 3 of Table II and Fig. 22(B).

Furthermore, there is a suggestive case in the z-plane which does not appear in the s-plane. When the root-locus of a sampled-data system passes through the origin, it is possible to cause the system to have a finite settling time. If the gain is equal to the value of K at the origin of the z-plane, the input-output z-transform becomes a

Eliahu I. Jury: It is true, as Mr. Mori points out, his interesting discussion in that the root locus can be plotted with the use of the other methods he mentions. However, for synthesis purposes and root-locus shaping it seems that the phase-angle loci approach is very promising, for it yields the form of the transfer function to be used for further stabilizing the system, as shown in Fig. 13.

The damping criterion indicated by Mr. Mori is useful in certain cases, but it yields only the complex roots. However, the system can also have the same overshoot if the roots are lying on the negative part of the real axis in the z-plane, as indicated in Table II. This part of the locus is not apparent in this logarithmic spiral, for Mr. Mori assumes only complex conjugate roots. Furthermore, this constant overshoot curve, which is limited to the sampling instants only, can be extended[1,2] for the actual response of the system, as shown in Fig. 23.

Fig. 22. Twenty-five-per-cent damping criterion on (A) the z-plane; (B) the s-plane

The locus of equation 61 is a logarithmic spiral as shown in Fig. 22(A). Of course, this spiral has infinite intersections with axes, and may have many intersections with the root locus. But practically, the wave form should not contain frequencies higher than 1/2 cycle per second, so the part of the
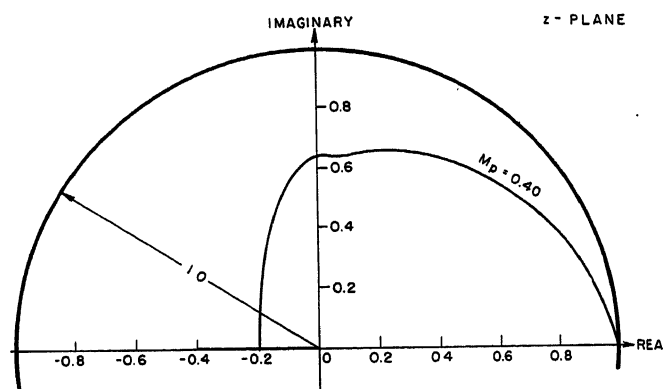
Fig. 23. Constant overshoot locus in the z-plane, for the following system

$$G(s) = \frac{\omega_n{}^2(1 - \epsilon^{-TS})}{s^2(s + 2\omega_n\eta)}, \quad M_p = C_{max} \ (n, m) \quad T - 1, \ \omega_n T, \eta \text{ are variables}$$

finite polynomial in z^{-1} as mentioned in reference 3. Some examples of the open-loop transfer functions in such case are $1/(z-1)$, $\epsilon^{-aTm}/(z - \epsilon^{-aT})$, $(z-1)/(z^2-z-1)$, etc.

I believe that the root-locus method has good applicability not only in continuous systems but also in sampled-data systems.

It is true that when the root locus passes through the origin, the system will have a finite settling time; but from critical study of these cases, it seems that often high overshoots appear between sampling instants, which are objectionable in most cases.

In conclusion, the author feels that the root-locus approach for synthesis of sampled-data control systems utilizing a digital computer or pulsed network is very promising. Further investigation of this method is indeed desirable.

REFERENCES

1. ROOT-LOCUS METHOD OF PULSE TRANSFER FUNCTION FOR THE SAMPLED-DATA CONTROL SYSTEMS, M. Mori. *Preprints for Automatic Control Symposium*, Japan, no. 61, 1954, pp. 5–17. Also *Automatic Control*, Japan, no. 1, 1955, pp. 5–12.

2. THE STUDY OF TRANSIENTS IN LINEAR FEEDBACK SYSTEMS BY CONFORMAL MAPPING AND THE ROOT LOCUS METHOD, V. C. M. Yeh. *Transactions*, American Society of Mechanical Engineers, vol. 76, 1954, pp. 349–61.

3. SAMPLED-DATA PROCESSING TECHNIQUES FOR FEEDBACK CONTROL SYSTEMS, Arthur R. Bergen, John R. Ragazzini. *AIEE Transactions*, vol. 73, pt. II, Nov. 1954, pp. 236–47.

REFERENCES

1. *Discussion* by Eliahu I. Jury of EXTENSION OF CONTINUOUS-DATA SYSTEM DESIGN TECHNIQUES TO SAMPLED-DATA CONTROL SYSTEMS, G. W. Johnson, O. P. Lindorff, C. G. A. Nordling. *AIEE Transactions*, vol. 74, pt. II, Sept. 1955, pp. 259–60.

2. SYNTHESIS AND CRITICAL STUDY OF SAMPLED-DATA CONTROL SYSTEMS, E. I. Jury. *Report*, University of California, Berkeley, Calif., series 60, issue no. 136, 1955.

Simplified Test Method for the D-Axis Transient Reactance and Time Constant

C. F. YOHE
ASSOCIATE MEMBER AIEE

A RELATIVELY simple test method is demonstrated to obtain the transient reactance and time constant of an a-c synchronous generator by an analysis of an oscillogram of the a-c generator field current. This method is applicable in many aircraft electric system studies, such as voltage-regulating stability and voltage rise on sudden removal of load, where only the a-c generator transient effects need to be considered. The influence of the d-c and subtransient armature currents are disregarded because of their comparatively negligible effect on the magnitude of the transient voltage of aircraft generators. Hence, to obtain test values of machine constants for these studies, the more efficient simplified test method may be used.

The subtransient effects are not reflected in the main field current, hence the direct-axis subtransient reactance X_d'' and the time constant T_a'' cannot be determined by this simplified test method. For those cases, where it is necessary to determine all of the direct-axis short-circuit constants, the test must be taken by the more general method given in reference 1.

Simplified Test Method to Determine X_d' and T_d'

TEST PROCEDURE

To determine the a-c generator direct-axis short-circuit transient reactance and time constant by test, the a-c generator is connected as shown in Fig. 1. The generator is operated at conditions of rated frequency, constant field excitation, and no load. The terminal voltage should be approximately on the air-gap line to avoid appreciable saturation. An oscillogram, as shown in Fig. 2, is taken when a short circuit is suddenly applied.

DETERMINATION OF X_d' AND T_d'

The symmetrical armature currents (sustained, transient, and subtransient) produce a rotating magnetomotive force (mmf) which is stationary with respect to the field poles while the unidirectional component produces an mmf which is stationary in space. Each of the arma-ture-current components has an associated component of current in the rotor circuits in the following manner:[2,3]

1. The sustained component of the short-circuit armature current is proportional to the sustained constant field-excitation current.

2. The direct-axis transient component of the armature current is proportional to the transient component of the field current.

3. The d-c component of the armature current is associated with a fundamental a-c component in the main field circuit and damper windings.

4. The subtransient component of the armature current is associated with a d-c component of current in the damper windings.

The presence of these related main-field circuit components to the armature short-circuit currents can be seen in the sample oscillogram shown in Fig. 2.

There is at all times a proportionality between the direct-axis transient component of the armature current and the transient component of the main-field current. Thus, by multiplying the transient component of the field current by the appropriate factor K, this curve may be plotted in terms of per-unit (pu) armature current, as shown in Fig. 3. The conversion factor K is determined by dividing the sustained pu short-circuit armature current, which is metered by an indicating instrument at the time of test, by the units of deflection of the sustained field current measured on the oscillogram. Thus

$$K = \frac{I_s \text{ (pu sustained short-circuit armature current)}}{I_{fs} \text{ (sustained field current in units of deflection)}}$$

The transient component of field current is defined here as the instantaneous value of field current (neglecting the a-c

component) minus the sustained current; see Fig. 2. The initial value of the transient component of armature current is determined by extrapolation of the curve, shown in Fig. 3, to the time of application of the short circuit. The transient reactance is determined by dividing the open-circuit voltage by the sustained component of the armature current plus the initial value of the transient component of the current obtained from Fig. 3. The transient time constant is determined as the time required for the transient component to decay to 36.8 per cent of its initial value.

ARMATURE SHORT-CIRCUIT TIME CONSTANT T_a'

This may be determined from an analysis of the a-c component of the field current. The d-c component of the armature current produces a mmf which is stationary in space. This mmf induces a fundamental a-c component of current in the main field circuit which decays at the same rate as the d-c armature-current component.[4] Thus, T_a' may be determined from the decrement factor of the a-c component of field current.

This can be done graphically by drawing a smooth curve through the peaks of the a-c component of field current and plotting the magnitude of this envelope on semilog graph paper as a function of time. T_a' is the time required for the envelope to decay to 36.8 per cent of its initial value.

In general, the d-c armature-current component does not persist after the first 2 or 3 cycles in aircraft synchronous generators. Hence, enough points are not available to determine accurately T_a' in these cases.

EFFECTIVE FIELD TIME CONSTANT T_{de}'

The transient component of the field current following load disturbances provides a convenient method of determining T_{de}' at any operating point. This time constant may be determined from the transient component of the field current following a step change in
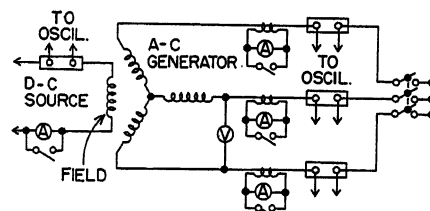
Fig. 1. Test setup for the determination of the a-c generator short-circuit constants
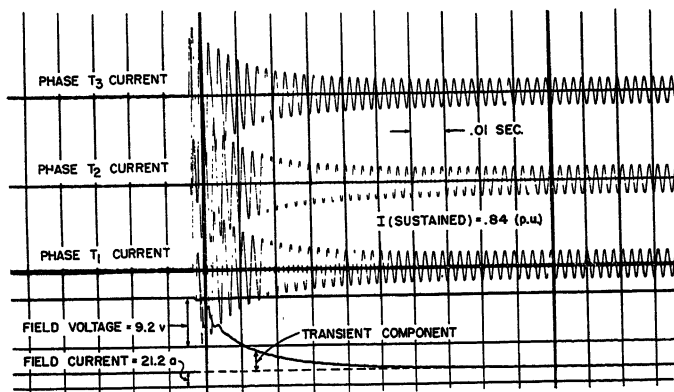
Fig. 2 (left). Oscillogram of 3-phase short-circuit currents

Fig. 3 (right). Semilog plot of the transient component of field current to determine X_d' and T_d'
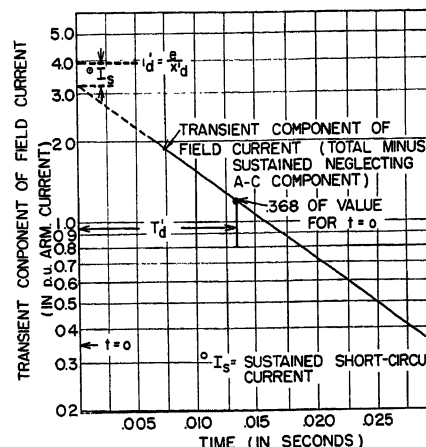
load. If the a-c generator is saturated at this operating point, the transient component of the field current will only approximate an exponential function of time and, therefore, will not yield a single-valued time constant.

Advantages of the Simplified Method

In applying the AIEE test method,[1] an oscillogram of the 3-phase short-circuit armature currents is required. To obtain such an oscillogram, several repetitions of the test may be required before proper oscillograph adjustments can be determined. For example, to measure an a-c generator transient-time constant in the order of 12 to 15 milliseconds requires oscillograph chart speeds of 150 inches per second to allow the transient component of current to decay to 36.8 per cent of its initial value in approximately 2 inches of film. This fast chart speed accentuates the problem of obtaining an oscillographic record of the armature currents which have large deflections and clearly defined traces. Several attempts are required before proper adjustments of oscillograph sensitivity, light intensity, chart speed, trace spacing, and optimum developing procedure is determined.

If the simplified method is applied, the foregoing difficulties become less critical. The oscillograph writing time is improved

since the field-current trace is a d-c transient as compared to the 400-cycle-per-second armature-current transient. Thus, light intensity adjustment, trace spacing, and developing procedure become lesser problems. The analyzing procedure is simplified since only one d-c exponential function is analyzed as compared to the three more complex a-c short-circuit current waves.

Limitation of Field Current Method

SUBTRANSIENT CONSTANTS

The subtransient effects are not reflected to the main field circuit, hence these constants X_d'' and T_d'' cannot be determined by this simplified test method. They must be determined by the more general method given in the AIEE test code.[1] However, if in addition to X_d' and T_d' only X_d'' is required, the latter may be determined by the static impedance method given as method no. 3 in the test code, while the former may still be obtained by the simplified test method.

SATURATION

If saturation exists, the proportionality between the armature short-circuit current and field current is only approximate. To minimize the effect of saturation,

the open-circuit voltage before short circuit should be approximately on the air-gap line.

SECOND-HARMONIC COMPONENT OF ARMATURE CURRENT

There is also a second-harmonic component of the armature current present in the armature short-circuit current wave. Associated with this is a d-c component of current in the field circuit which has a decrement factor of T_a'. This d-c component of field current will disrupt the proportionality between the armature current and the field current. For most aircraft synchronous generators this effect does not have an appreciable influence after the first 2 or 3 cycles. Thus, this influence does not usually hinder the determination of the transient constants by the simplified method.

VERIFICATION OF TEST METHOD

Table I shows the data obtained by analyzing oscillograms by both the method given in reference 1 and the suggested method given herein. There is less than a 5-per-cent discrepancy between the data obtained by each of the methods. This discrepancy can be contributed to saturation at rated voltage and to inherent errors of the graphical analysis of the oscillograms.

Conclusions

The suggested test for determining the transient reactance X_d' and the transient time constant T_d' simplifies both the test procedure and method of analyzing the oscillogram. The constants obtained on aircraft generators by the simplified method compare favorably to the constants determined by the method given

Table I. Comparison of Data Obtained by Analysis of Armature-Current Wave and Field Current

Machine	AIEE Test Code No. 503[1]		Simplified Method	
	X_d', Pu	T_d', Seconds	X_d', Pu	T_d', Seconds
40-kva aircraft, 360/440-cycle-per-second salient-pole generator	0.217	0.0122	0.219	0.0139
60-kva aircraft, 360/440-cycle-per-second salient-pole generator	0.282	0.0127	0.270	0.0133
	for rated voltage and 400 cycles per second			

in the AIEE test code.[1] This test can also be used to determine the effective field time constant $T_{dz}{}'$.

Consideration should be given to the effects of saturation and second-harmonic component of the armature current. The error caused by saturation should be minimized by testing at an unsaturated condition. The effect of the second-harmonic component of armature current will be negligible for those generators where $X_d{}''$ and $X_q{}''$ (quadrature-axis

subtransient reactance) are nearly equal[5] and/or $T_a{}'$ is considerably smaller than $T_d{}'$.

This simplified test method will expedite the time and effort required to determine $X_d{}'$ and $T_d{}'$ for those transient studies where the subtransient effects are to be neglected. This test can also serve as a convenient method to verify the constants determined by calculations as well as those determined by the AIEE test method.

References

1. TEST CODE FOR SYNCHRONOUS MACHINES. AIEE Standard No. 503, June 1945, pp. 32–40.

2. THREE-PHASE SHORT-CIRCUIT SYNCHRONOUS MACHINES—V, R. E. Doherty, C. A. Nickle. AIEE Transactions, vol. 49, Apr. 1930, pp. 700–14.

3. ELECTRICAL TRANSMISSION AND DISTRIBUTION REFERENCE BOOK. Westinghouse Electric Corporation, East Pittsburgh, Pa., 4th ed., 1950, pp. 153–55.

4. Doherty and Nickle. Op. cit., p. 703.

5. ANALYSIS OF SHORT-CIRCUIT OSCILLOGRAMS, W. W. Kuyper. AIEE Transactions (Electrical Engineering), vol. 60, Apr. 1941, pp. 151–53.

The Development of a Static Voltage Regulator for Aircraft A-C Generators

H. H. BRITTEN
ASSOCIATE MEMBER AIEE

D. L. PLETTE
ASSOCIATE MEMBER AIEE

THE purpose of this paper is to describe the development of a static aircraft voltage regulator for constant- and variable-speed a-c generators. This type of regulator is one which utilizes components which are static in nature as a means of control. The development was started in an effort to develop a new and better regulator to meet more rigid specification requirements. The regulator that has resulted not only meets these requirements but promises inherent reliability and long life.

Background

INCREASED USE OF ELECTRIC POWER IN AIRPLANES

In modern aircraft there is a continuous trend toward increased use of electrical power.[1] This trend is noticeable in most aircraft but more pronounced in combat planes. Realization that functions can be performed electrically with greater reliability and economy of weight has increased the burden on generators and control. Hydraulic systems, for example, are more vulnerable to battle damage than electrical systems. Increased use of larger and faster-moving turrets places additional demands on the electrical system. That planes may be effective in all weather conditions, radar and other types of electrical equipment are considered essential today. These factors plus many others are increasing the bur-

den on the aircraft electrical system and bring about greater dependency of the modern plane upon its electrical system.

TREND TOWARD ALTERNATING CURRENT FOR AIRPLANES

During and after World War I, the development of d-c systems generally followed the practice of the automotive industry. The 24-volt battery and the 28-volt d-c generator became the standard before the beginning of World War II. By this time, important functions were being performed electrically, and the safety of the plane depended upon the reliability of the electrical system.

As electrical loads and aircraft sizes increased, it became apparent that a more economical generating and distributing system was required.[1] Higher voltages were needed to reduce conductor weights. Two methods of achieving this were selected as having the most promise. These were the 120/208-volt 3-phase a-c system, and the 120-volt d-c system. Because of certain advantages inherent in the a-c system, there is a strong trend in this direction. This trend has been further accelerated by the development of practical constant-speed drives.

Some of the advantages attributed to the a-c system are:

1. Ease of obtaining any desired utilization voltage by means of a transformer.

2. Decreased brush wear and decreased commutation problems.

3. Simplicity and reliability of induction motors.

The importance of this trend toward a-c power was demonstrated by the World War II B-36 aircraft which has an a-c generating capacity of 160 kva. Much of this power was used as alternating current and the d-c power was obtained through transformer-rectifier units.

It has been stated that the dependability of an aircraft generating and distributing system should compare favorably with that of an electric utility, with the exception of ultimate life expectancy.[1] The dependability of such a system cannot exceed the reliability of the components which make up the system. One of these system components is the generator voltage regulator.

IMPORTANCE OF VOLTAGE REGULATOR

The voltage regulator is of prime importance in the successful operation of an electrical system. Good voltage control under all conditions is essential as many load devices require substantially normal voltage for satisfactory operation and life. With increased dependence of the plane upon its electrical system, regulator specifications have become steadily more rigid in their requirements. It was therefore believed, that the best solution of the over-all problem would be to develop a regulator using static components which would offer performance, life, and reliability commensurate with the increasingly important role of electric power in the modern aircraft.
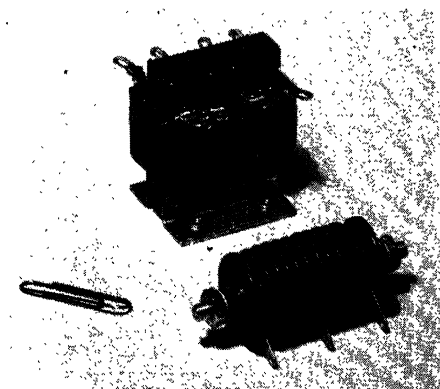
Fig. 1. Magnetic amplifier for aircraft use, consisting of saturable reactor and associated rectifier. Rated 5 watts' continuous output at 320 cps

Basic Considerations

In the initial stages of a regulator development, basic means of accomplishing various functions are investigated in order to determine that method which affords the greatest reliability, best performance, and lightest over-all regulator weight.

TYPE OF AMPLIFICATION AND CONTROL

It was decided that the magnetic amplifier should be used as the primary means of amplification and control. A magnetic amplifier is any device which utilizes the saturable reactor as a means of control or power amplification. The particular type of magnetic amplifier which was applied to this regulator is the amplistat.[2] This magnetic amplifier is

characterized by its ability to obtain extremely high power gain and speed of response. Its output is essentially direct current and hence can be used to control one or more fields on the exciter.

The saturable reactor and its associated metallic rectifier are extremely rugged and dependable when properly applied. At frequencies close to 400 cycles per second (cps) the physical size of this type of magnetic amplifier per watt of output is small enough to be practical for aircraft use. Fig. 1 shows a magnetic amplifier capable of 5 watts' continuous output at 320 cps.

CONTROL OF EXCITER

Conventional aircraft a-c generators with integral exciters are controlled by varying the excitation to one or more fields on the exciter. It was decided that this regulator would control the exciter by supplying positive ampere-turns to the exciter field in varying amounts depending upon the condition of regulation. This method of exciter control was chosen because it required minimum power from the regulator for control.

In variable frequency a-c systems of this type, the two extremes of control requirements are found at: 1. full generator load, minimum rated speed, and 2. zero generator load, maximum rated speed. In addition to these steady-state requirements for control, additional range of control must be provided to take care of short-time overloads and insure proper system transient performance in recovering from sudden load changes. The ef-
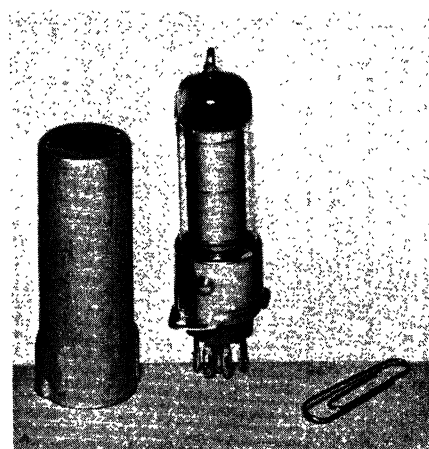
Fig. 2. Glow tube voltage reference

fects of ambient temperature and altitude on generator excitation requirements must also be considered in determining the full range of control which is necessary.

The maximum amount of power necessary for control can be reduced by the utilization of more exciter field winding space. This process is limited by the total amount of space available and also by the fact that using more field space causes a longer exciter field time constant. Since the latter is undesirable from a response and stability standpoint, some compromise must be reached whereby satisfactory performance can be achieved without adding unnecessary weight to the regulator.

TYPE OF REFERENCE

The selection of a reference for a voltage regulator is of prime importance since in any automatic control system the ability of a regulator to control a quantity depends on the comparison of that quantity with a known or fixed quantity. It is generally desirable to reproduce the relatively fixed properties of the reference quantity in the output quantity. Therefore, any drift or other instability which may be found in the voltage reference will
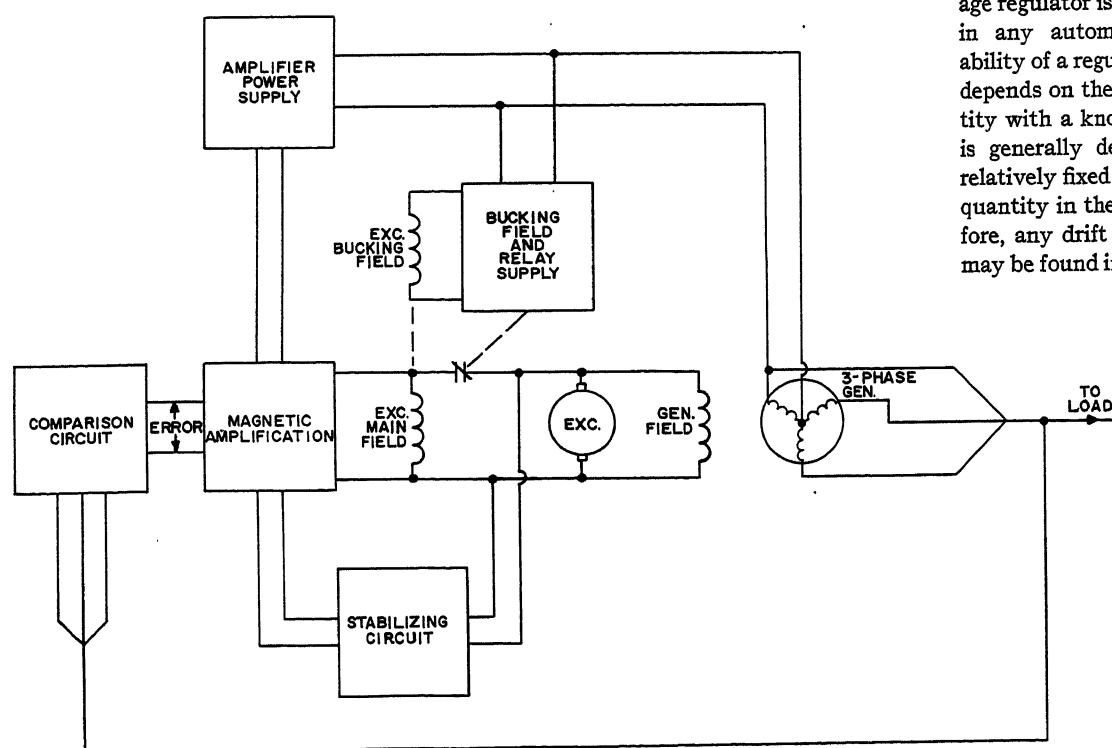
Fig. 3. Regulator block diagram

Fig. 4. Comparison circuit

Fig. 4. Comparison circuit

E_R = REFERENCE VOLTAGE
E_{FB} = FEEDBACK VOLTAGE
E = ERROR SIGNAL TO AMPLIFIER

3-PHASE GENERATOR VOLTAGE

FULL-WAVE BRIDGE RECTIFIER

TEMPERATURE COMPENSATION

GLOW TUBE REFERENCE
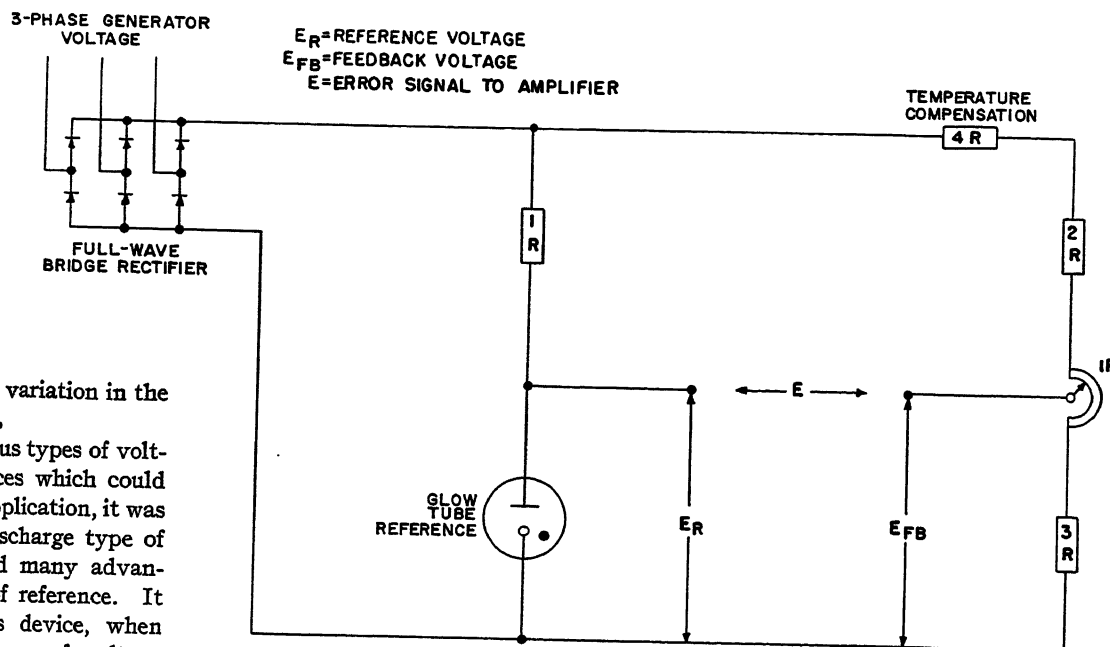
appear as an undesirable variation in the generator output voltage.

After considering various types of voltage and current references which could have been used for this application, it was decided that the glow-discharge type of tube (glow tube) showed many advantages over other types of reference. It is well known that this device, when supplied from a d-c source of voltage through a series-dropping resistor, will maintain a nearly constant voltage across its terminals. This action is relatively independent of variations in the magnitude of the voltage source and hence this reference has also become known as a voltage regulator tube.

The glow tube in its miniature form weighs only $1\frac{1}{2}$ ounces including the weight of the mounting socket and protective shield. The unit is approximately $3\frac{1}{4}$ inches long and 7/8 inch in diameter, and so takes up very little of the available regulator space. Fig. 2 shows this reference tube.

The voltage which the glow tube maintains across its terminals will vary less than 1/2 per cent (%) when the ambient temperature is varied from −55 to +71 degrees centigrade. Since it operates from a direct voltage, the glow tube output voltage is not affected by changes in frequency.

In spite of the fact that this reference is an electronic tube, it is an extremely rugged device. Unlike most tubes, it has no filaments and no grids. Extended life tests under vibration cycling at various frequencies have been performed on groups of these tubes. The tests have shown that this type of unit is capable of withstanding vibration tests 10 times the duration of those called for in most of the existing regulator specifications without causing malfunction of the regulator.

In addition to these advantages, the glow tube has sufficient reference power available to compare favorably with other references considered, whose weights were many times greater. The importance of the amount of reference power will be considered later.

Specific Design Considerations

DESCRIPTION OF VOLTAGE REGULATOR SYSTEM

The voltage regulating system consists of the static-voltage regulator, rotating d-c exciter, and a-c generator. The elements of the system may be represented by functional blocks as shown in the block diagram of Fig. 3. The comparison circuit accomplishes the voltage sensing and contains such elements as a glow tube, dry-disk rectifiers, and resistors.

The output of the comparison circuit is

AMP.-TURNS AT_A

AT_C = FIRST STAGE CONTROL AMP. TURNS.

AT_A = SECOND STAGE CONTROL AMP. TURNS.
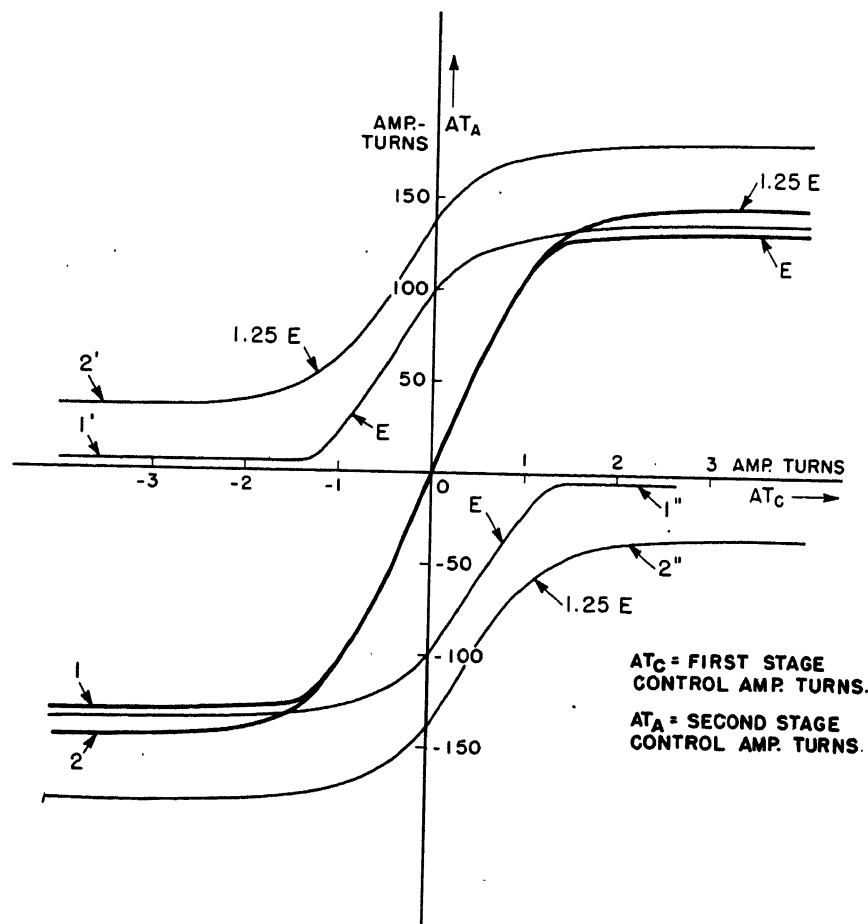
Fig. 5. Effects of supply voltage variation on push-pull amplifier characteristic
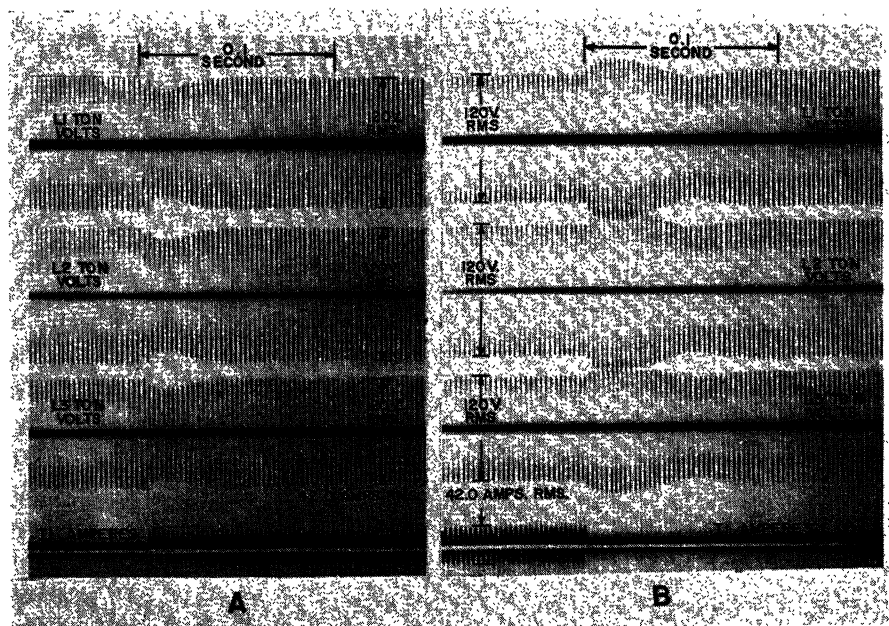
Fig. 6. Application and removal of full load

A—Application of full load, 0.75 pf. At 6,000 rpm, 15-kva generator is controlled by static voltage regulator

B—Removal of full load, 0.75 pf. At 6,000 rpm, 15-kva generator is controlled by static regulator

a voltage directly proportional to the difference between the generator voltage and the regulator set voltage and is referred to as the error signal. If the characteristics of the error signal are preserved with respect to linearity and its power level increased sufficiently, it can control the excitation of the a-c generator and cause its output voltage to remain relatively constant regardless of load, speed, and power factor (pf).

Two stages of magnetic amplification are used to increase the power level of the error signal. These are referred to on the block diagram of Fig. 3 as magnetic amplification. Included in this block are such elements as saturable reactors, rectifiers, linear reactors, and resistors. The stabilizing circuit contains such elements as capacitors and resistors. Its function will be explained later.

The action of the voltage-regulating system can be traced simply by assuming some steady-state condition and then considering that for some reason the generator voltage tends to decrease. This causes an error signal from the comparison circuit in such a direction as to increase the output of the magnetic amplifier. This action increases the output of the d-c exciter, which tends to restore the generator voltage to its original value.

THE COMPARISON CIRCUIT

The comparison circuit in a voltage regulator usually consists of two parts;

a voltage reference and a directly proportional feedback from the generator output voltage which it is desired to regulate. Fig. 4 shows the comparison circuit as used in the static regulator and includes a temperature-compensating resistor. The need for temperature compensation will be discussed later.

As shown in Fig. 4, the output of the 3-phase a-c generator voltage is passed through a full-wave bridge rectifier to obtain a direct voltage to supply the comparison circuit. The output of this rectifier will be nearly proportional to the average of the three line-to-line voltages of the 3-phase a-c generator. Therefore if a fixed proportion of this voltage is compared to the constant voltage present across the glow tube, a means of telling whether the generator voltage is too high or too low is achieved. The difference between the reference voltage and this direct feedback voltage is a measure of the amount and direction of the generator voltage deviation from that voltage which the regulator is set to maintain. This difference or error signal is the output of the comparison circuit and is used to control the magnetic amplifier.

As explained previously, one of the prerequisites of a good reference is its power output. This property is important because it directly influences the available power in the error signal from the comparison circuit. Referring to the comparison circuit of Fig. 4 and specifically to the glow tube references, the

maximum reference power available Pr may be defined as the product of the glow tube voltage and its maximum rated current. Furthermore, the power output of the comparison circuit Po may be defined as the power absorbed by the resistance of the control circuit of the first-stage magnetic amplifier. Assuming that equal bidirectional linearity is desired in the error signal, it may be defined as a condition where the maximum absolute value of current in the error signal at the end of linearity is equal in both positive and negative directions. The end of the linearity is assumed to occur when the current in the glow tube falls to minimum current or reaches the maximum current for which it was designed. Under the foregoing assumptions the maximum power that can be obtained from the comparison circuit may be expressed very closely by

$$Po = \frac{\alpha^2 Pr}{200 \alpha_M}$$

where

Po = power output of comparison circuit
Eo = normal generator voltage at which voltage Po = 0
Em = generator voltage at which glow tube conducts maximum rated current
α = generator voltage deviation from Eo in per cent
Pr = maximum reference power available
α_M = per cent of voltage over which output of comparison circuit is linear with respect to voltage = $(Em - Eo)/Eo \times 100\%$

Often, transient conditions can occur which limit the amount of power that can be obtained from the comparison circuit for small voltage deviations. If, for example, the reference capabilities were extended in a relatively narrow voltage deviation from normal voltage, the reference may be damaged during a severe voltage transient as experienced when clearing a 3-phase short circuit on the generator. Nevertheless, in general, the amount of power available in the error signal is proportional to the reference power available and may be said to be tempered by the maximum reference power transiently permissible.

AMPLIFIER POWER SUPPLY

Any power amplifier requires an auxiliary power supply of some kind. The magnetic amplifier operates from an a-c supply. The logical place to obtain this power is directly from the a-c generator since there would then be no dependence on external sources of power outside of the generating system itself. It was therefore decided that the amplifier
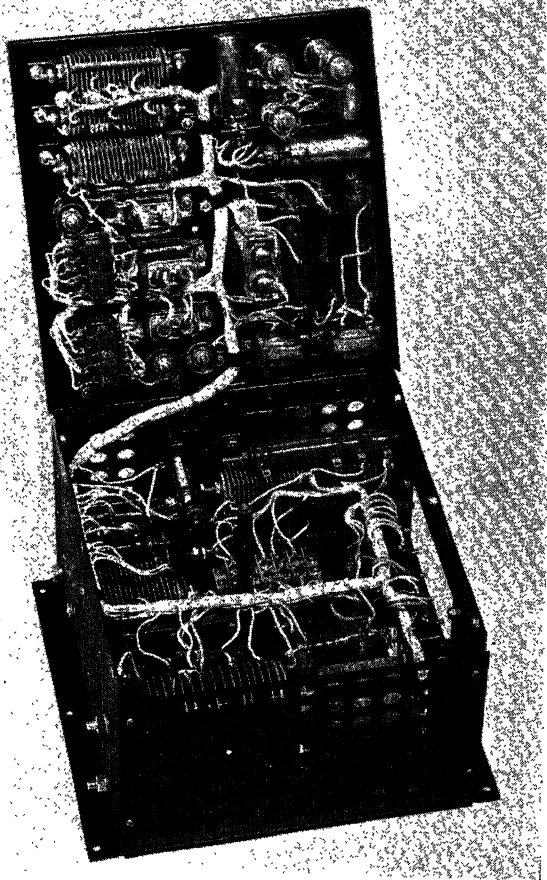
should be supplied from one single-phase voltage of the generator. The power thus supplied is passed through a transformer in the regulator to provide isolation from the generator and to deliver correct utilization voltages.

Obtaining the amplifier power supply from the generator requires special consideration. There are times when excitation is required and no voltage is available to supply the amplifiers. Such conditions exist during:

1. Initial build up of system voltage from rest.

2. Three-phase short circuit or a single-phase short circuit on the generator which eliminates the amplifier power supply. Under either of these conditions, it is essential that very large values of excitation be attained in order to cause large generator armature current to flow for rapid clearance of the existing short circuit.

These two obstacles are overcome by inserting a small single-pole, single-throw, normally closed relay contact between the exciter field and the exciter armature; see Fig. 3.

With the relay in the closed position, the exciter is self-excited. If the coil of this relay is supplied from the same source of voltage as the amplifier, then the

excitation on the generator will build up until the amplifier has sufficient power available to supply the exciter field excitation. At this point, the relay will pick up and the exciter will no longer be self-excited. Since this relay pickup will not occur in the case of the two types of short circuits mentioned, the exciter will provide ceiling voltage and thereby aid in clearing the short circuit from the system.

Magnetic Amplifier Design

In general, power gain and time constant are interdependent in magnetic amplifiers. This relationship may be expressed approximately as

$$P_G = KT_c$$

P_G = power gain of the magnetic amplifiers
T_c = time constant of the amplifier
K = a constant dependent on the amplifier size, supply voltage frequency, characteristics of the core material, etc.

It now becomes apparent why the power available in the error signal is important. The power level of the error signal is to be increased by magnetic amplification to the proper level to supply field power to the exciter. Therefore, the amount of power gain required by this amplifier can be reduced by increasing the power available in the error

signal. As seen in the foregoing equation, this will result in a smaller time constant for the amplifier, which is very desirable from a response and stability standpoint.

In the static regulator a 2-stage amplifier is used to obtain the required power gain with extremely fast response time. The advantage of a 2-stage amplifier can readily be seen by a simple example: Assume that a power gain of 90,000 is required. If $K = 90,000$ in the foregoing expression, a time constant of 1 second would be required in a 1-stage amplifier. If a 2-stage amplifier were used, each stage having a power gain of 300 and $K = 90,000$, the necessary power gain would be achieved and each amplifier would have a time constant of 0.0033 second.

The first stage of magnetic amplification in the static regulator consists of two magnetic amplifiers connected in a push-pull circuit. This type of connection tends to minimize the effects of varying voltage and frequency of the amplifier power supply on the over-all transfer characteristic. The output of each amplifier in the push-pull circuit is fed into separate control windings on the second-stage amplifier. These control windings are connected in opposition and the difference in ampere-turns in the two windings is the resultant control ampere-turns in the second-stage amplifier. The curves of Fig. 5 show the individual and resultant characteristics of a typical push-pull stage feeding into a second-stage amplifier.

Curves 1' and 1'' show the individual first-stage characteristics and curve 1 the resultant push-pull characteristic at normal voltage. Curves 2', 2'', and 2 show the same characteristics at 25% overvoltage. These curves clearly illustrate the minimized voltage effects on this type of circuit. In a similar manner the push-pull circuit tends to minimize the effects caused by frequency variation of the amplifier power supply.

The complete coverage of all problems encountered during the development with respect to magnetic-amplifier design are beyond the scope of this paper. Severe voltage transients and wide frequency range of the amplifier power supply altered the design procedure from that used on amplifiers power supply altered the design procedure from that used on amplifiers which are to operate from fixed frequency and supply voltage. In addition to this, extensive investigation was made to determine the proper core materials for achieving maximum performance and minimum weight.

Exciter Bucking Field

On variable-frequency a-c generators, the range of excitation requirements can be very great. On some generators, for example, the exciter field power required at full load and minimum speed might be 400 times the power necessary to provide the same voltage at zero load and maximum speed. By addition of a few negative or bucking ampere-turns to the exciter, the power requirements at top speed and no load can be multiplied by 10. These same ampere-turns require less than 25% additional exciter field power at the minimum speed and full-load condition. Therefore the magnetic amplifier has to supply only a 50-to-1 power range, a job which is considerably easier. For this reason, the exciter is supplied with a bucking field which receives its excitation from rectified generator voltage.

A further advantage of this bucking field is that it decreases the time necessary to return to rated voltage from the high voltages present after removal of heavy loads and short circuits at high generator speeds.

Stabilizing Circuits

In general, feedback control systems having more than two significant time constants and having relatively high gain are capable of producing sustained oscillations, sometimes referred to as "hunting." To eliminate this condition, the static regulator employs the use of a capacitor-resistor network in the feedback circuit from the exciter. It is sufficient to state that this circuit provides such necessary phase margin that the regulating system is stable and sufficiently damped to meet the transient requirements.

Temperature Compensation

To avoid variations in regulated voltage due to the effects of changing temperature on the characteristics of various components, some means of temperature compensation must be included in the regulator. This is provided by a special temperature-compensating resistor which is inserted in the comparison circuit, as shown in Fig. 4. The large change in the resistance of this element with temperature tends to compensate for changes in other elements so that nearly constant voltage is held, regardless of ambient temperature.

Parallel Operation

Regulators designed to operate with generators which are to be paralleled must have some provision for reactive load equalization between generators. Many alternating voltage regulator supplied today accomplish this by placing a small current transformer in series with one of the generator lines. This current transformer feeds an equalizing reactor transformer, the output voltage of which is shifted approximately 90 degrees in phase relation to the current from the current transformer. When this voltage is added to the generator line voltages in the proper manner at the input to the regulator, the action of the regulator is such as to reduce excitation of the generator if the generator is producing more than its share of excessive lagging or overexcited reactive current. Leading or underexcited reactive current from the generator would cause the regulator to increase the generator excitation by action of the equalizing circuit. The real or power component of current has little effect on the operation of this circuit. Division of real power is accomplished by governors on the generator drives.

A further refinement of the equalization circuit is obtained by connecting the secondaries of the current transformers of all paralleled generators in a series loop arrangement. By this connection a signal proportional to the difference in the reactive currents carried by the various generators is supplied to the regulator-sensing circuits. This has become a rather standard feature of aircraft a-c generator systems and is known as differential equalization. The advantage of this type of equalization is that a relatively high gain may be employed in the reactive load equalization circuit without causing abnormal droop in the system voltage when reactive loads are impressed on the system.

Results

The following are some typical data taken with a 15-kva 0.75-pf, 4,800 to 7,200-rpm generator and the static aircraft voltage regulator.

1. Voltage variations due to load changes from no load to full load and speed changes from 4,800 to 8,000 rpm were held to less than ±1%.

2. Voltage variation due to changes in regulation ambient temperature from −55 to +71 degrees centigrade was less than ±1/2%.

3. Recovery time after application and removal of full load was 0.05 second at minimum rated speed of 4,800 rpm and 0.1 second at maximum rated speed of 7,200 rpm. In Fig. 6 are oscillograms showing the recovery time obtained with this regulator and the 15-kva generators.

4. Output voltage modulation under worst conditions was less than 0.5%.

5. Voltage was unaffected by operating position, vibration, or an acceleration of 10 g along any axis of the regulator.

The weight of the regulator was approximately $17^1/_2$ pounds and its size was approximately 11 by 11 by 7 inches. The finished regulator is shown in Fig. 7.

Conclusions

As a result of this development, two regulators have been designed which are nearly identical. One of these units is capable of operation with 15-, 30-, 60-, and 90-kva generators which have an operating speed range of 4,800 to 7,200 rpm, (320 to 480 cps). The other unit is capable of operation with 15- and 30-kva generators with speed range from 4,000 to 8,000 rpm, (400 to 800 cps).

A number of regulators have been built for each rating of generator, and tests have shown them to have excellent electrical performance even under the most adverse conditions of temperature, load, acceleration, and vibration. In addition, these regulators promise inherent long life and low maintenance as well as the ruggedness and dependability so vital to the success of our modern aircraft.

References

1. Basic Design Principles for A-C Electric-Power Systems in Large Military Aircraft, C. K. Chappuis, Leonard M. Olmsted. *AIEE Transactions (Electrical Engineering)*, vol. 65, Jan. 1946, pp. 12–17.

2. The Amplistat and Its Applications, H. M. Ogle. *General Electric Review*, Schenectady, N. Y., vol. 53, Feb., p. 32-ET; Aug., p. 41-RR and Oct., 1950 p. 41-IRY.

3. The Extension of Amplistat Performance by A-C Components, R. E. Morgan, H. M. Ogle, V. J. Wattenberger. *AIEE Transactions*, vol. 69, pt. II, 1950, pp. 986–91.

4. An Amplistat Regulator Applied to a 5-Kva 400-Cycle Aircraft Inverter, R. E. Morgan, J. A. Walley. *Ibid.*, pp. 1243–48.

Precision Low-Frequency Inverter for Aircraft

M. E. DOUGLASS
ASSOCIATE MEMBER AIEE

J. R. LAVENDER
NONMEMBER AIEE

THE purpose of this paper is to detail the power supply requirements for a mass rate fluid flowmeter; to examine various methods of driving the impeller of the flowmeter; and to describe the power supply for the drive method selected with its performance characteristics.

The power supply used is a 3-bar commutator, chopping direct current to provide a simulated 3-phase 4-cycle electrical output, that drives the impeller motor at 240 rpm. The commutator is driven at constant speed by a permanent magnet d-c motor slaved to a balance wheel. The frequency is maintained constant within 0.3 per cent (%) over the temperature range from -55 to $+71$ degrees centigrade in all positions; with vibration from 100 to 500 cycles per second (cps) at 10 g acceleration. The power output is approximately 40 watts and is free from radio noise in accordance with Military Specification *MIL-I-6181B*.[1]

Need for Constant Speed

In the mass rate flowmeter for which this power supply was developed,[2,3] fluid enters a unit, called an impeller, which is a cylinder with straight axial passages, sufficiently long so that the fluid flow at exit is parallel to the axis, and which is rotated about its axis at a constant speed. Immediately downstream from the impeller is a similar unit mounted in bearings so it is free to rotate but restrained from rotation by a spring; see Fig. 1. This unit, called the turbine, absorbs the energy from the fluid imparted by the impeller, and the torque developed in the turbine is proportional to the mass rate of flow of fluid and to the angular velocity of the impeller. This torque deflects the turbine against its restraining spring so the turbine

position which is used to indicate mass flow rate is proportional to the mass rate of flow times the impeller speed. Thus, any deviation in speed of the impeller is reflected directly as an error in the indication of the flowmeter. To achieve the high standards of accuracy required for the aircraft application of this flowmeter, deviations in speed of the impeller of only 0.3% from the base speed are permissible.

Miscellaneous Specifications

In addition to the constant speed specification, a further limitation that all electrical wiring must be sealed from the fluid passages was required for safety, and since the equipment was to be airborne, it must be of minimum weight, it must operate over a temperature range of -55 to $+71$ degrees centigrade, from sea level to 50,000 feet altitude, and over a voltage range of 22.5 to 30 volts for a 28-volt nominal system.

Possible Drive Methods

In the initial consideration of the problem, several angular speed regulating devices were investigated, and devices such as centrifugal switches, induction disk generators, and permanent magnet generators were considered as primary detectors for speed. Size, weight, reliability, accuracy, or development time required, narrowed the field to consideration of the three approaches pictured schematically in Fig. 2.

The first system of Fig. 2 used a regulated power supply, if it were available on the aircraft, to drive a synchronous motor. A magnetic coupling to provide sealing of the motor from the fluid was required in this system, and a gear train to reduce the speed of the synchronous motor to the 63 rpm required at the impeller. The major problem with respect to this proposal was that the aircraft power supplies available do not control frequency much closer than 2% to 5%, and while this is adequate for most purposes, it is about 10 times the limit allowable for the flowmeter application.

Where an accurately controlled frequency is required, it has been the practice to use an electronic power supply governed by a tuning fork or crystal. Application of this practice to the flowmeter problem provided systems 2 and 3 of Fig. 2. System 2 simply put a regulated power supply between the generator and the synchronous motor. Consideration was given to both 400-cycle and 60-cycle controlled frequency power supplies, of which there are quite a few commercially available to meet the required specifications. The 400-cycle systems showed a weight advantage over the 60-cycle systems; however, the weight of the regulated power supply was so large as to be a serious disadvantage.

The third scheme shown is the one finally selected. This one combines the motor and magnetic coupling in one unit. The windings of the motor are outside a sealed metallic wall, Fig. 3; a permanent magnet rotor running at 240 rpm is inside the wall in the fluid stream, and it drives the impeller through a gear train at 63 rpm. A 4-cycle frequency-regulated power supply provides the electrical energy to the motor.

The selection of 240 rpm as the motor speed provides an optimum balance between no-load losses of the motor, and torque required to drive the impeller. Since the required energy input to the impeller is fixed by the turbine torque output requirements, the higher the motor speed the less torque it needs to supply this energy, and the smaller it could be made. However, since the motor rotor runs in the fluid, the faster it runs, the higher are the no-load losses due to viscous drag, which would require extra energy from the power supply. With a 2-pole motor, 240 rpm requires 4 cps for synchronous operation, which establishes the frequency of the power supply.

The low-frequency power supply-motor combination provides some highly significant simplification and weight advantages over the conventional 400- or 60-cycle systems. In the flowmeter the low-speed motor with its unique design eliminated a magnetic coupling of about 1 pound estimated weight. This represents 6 to 8 pounds savings on the overall system on 6- or 8-engine planes. The gearing is simplified by using a 3.81 to 1 reduction instead of the 127 to 1 required for an 8,000-rpm, 400-cycle motor.

The power supply as finally developed provides about 37.3 watts (driving 8 flowmeters) with a weight of $2^1/_4$ pounds (i.e., 16.6 watts per pound) as compared
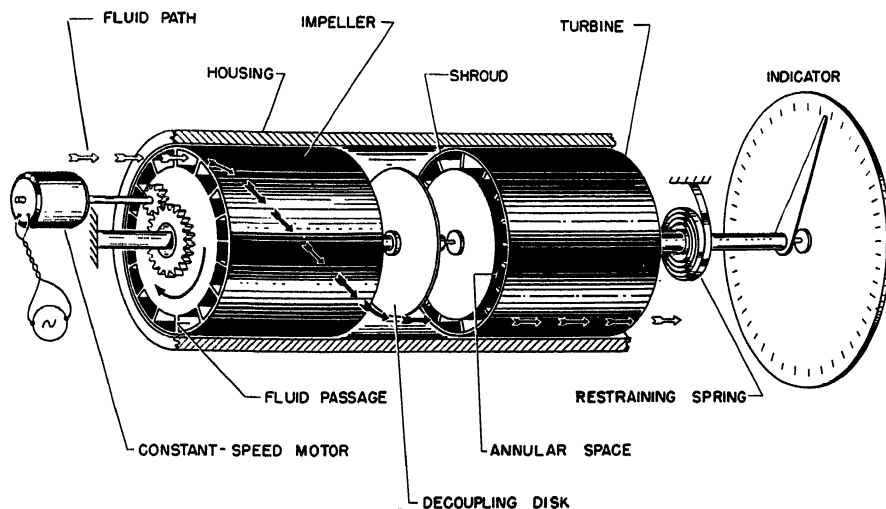
Fig. 1. Flowmeter operating principle

Fig. 2. Flowmeter drive systems

to an estimated weight of 21 pounds for a 400-cycle controlled frequency power supply to provide equivalent performance. (Part of this weight of the electronic power supply is required to supply the high no-load loss and high reactive volt amperes of 400-cycle synchronous motors.)

Power Supply

The power supply consists of a 3-bar commutator, each bar being connected electrically to a slip ring. Two brushes diametrically opposite on the commutator are connected to the positive and negative terminals of the 28 volt d-c power supply which is commonly available on most aircraft. Three brushes are in contact with the 3 slip rings and are connected to the coils of a 2-pole 3-phase Y-connected motor. As the commutator rotates, each of these 3 brushes is alternately positive, dead, negative, and dead. The resultant magnetic field in the motor stator rotates 30 degrees each time a switching operation takes place in the commutator. Fig. 4 shows a schematic diagram of the commutator. Table I indicates the switching sequence, the coil currents, and the resultant motor vector.

To rotate the commutator at constant speed, a small permanent magnet d-c motor whose speed is controlled by a balance wheel is used.[4] This motor is shown in Fig. 5. The mechanism is shown in Fig. 6. A pair of contacts is operated upon by a cam driven by the motor and a balance wheel, in such a manner that the motor receives sufficient energy from the power source to rotate a fixed amount during each swing of the balance wheel. Changes in motor load or voltage change the proportion of the

time the contacts are closed to maintain the armature rotation per cycle of the balance wheel constant. This provides a very accurate average speed control, and by the use of materials insensitive to temperature change in the balance wheel and hair spring, this accuracy is maintained over the required temperature range.

With this basic method accepted, the details of commutator design remained to be established. To select the time during the cycle at which the switching operations should take place, it was reasoned that if the low-order harmonics were eliminated from the current wave form, filtering (if found necessary) would be easier, and motor oscillation frequency, due to the power supply, would be raised well over the natural resonance of the turbine which is about 1 cps.

A Fourier analysis of the coil current, assuming only resistance in the motor coils, was made. This current has the
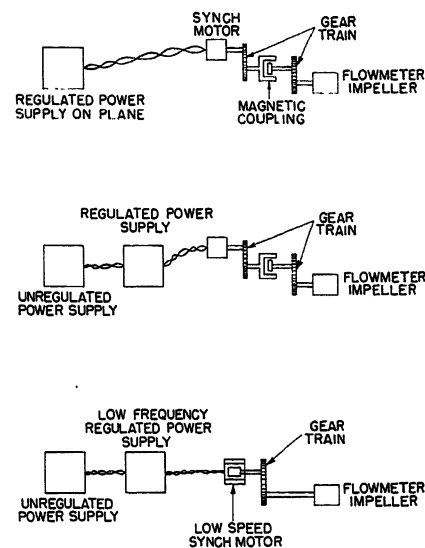
general form shown in Fig. 7. The relative magnitude of the vertical steps is established by the connections to the motor. The values α, β, and γ can be controlled, within limits, by the relative sizes of the commutator segments and brush width. From the symmetry of the wave form about the 180- and 90-degree ordinates, the Fourier series representing this wave is $f(\theta) = A/\sin \theta + A_3 \sin 3\theta + A_5 \sin 5\theta \dots A_m \sin m\theta$, where m has odd integral values. The coefficients of each term in this expression are given by

$$A_m = \frac{4}{\pi} \int_0^{\frac{\pi}{2}} f(\theta) \sin m\theta \, d\theta$$

Substituting proper values for $f(\theta)$ and integration limits

$$A_1 = \frac{2E}{3\pi R} (2 \cos \alpha + \cos \beta + \cos \gamma)$$
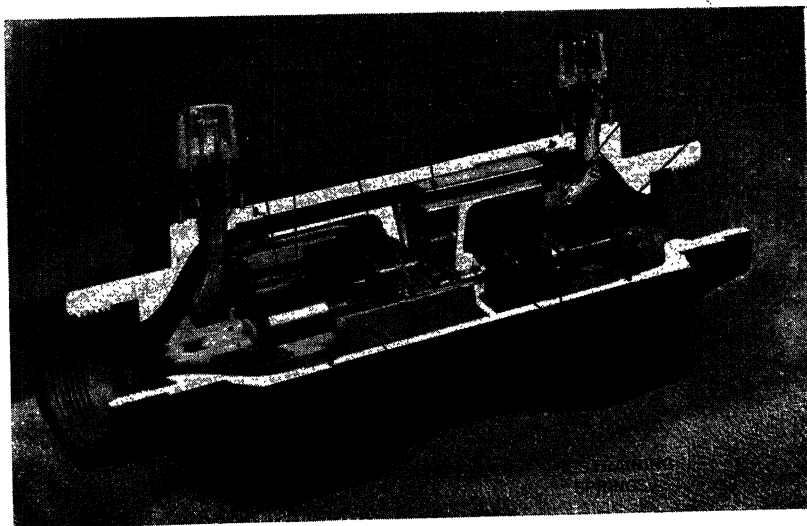
Fig. 3. Flowmeter cutaway
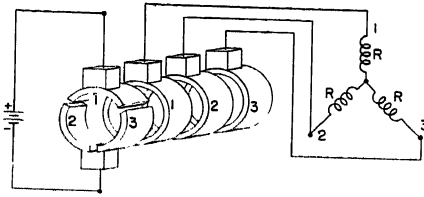
Fig. 4. Inverter and flowmeter motor operating principle

$$A_3 = \frac{2E}{9\pi R}\,(2\cos 3\alpha + \cos 3\beta + \cos 3\gamma)$$

$$A_5 = \frac{2E}{15\pi R}\,(2\cos 5\alpha + \cos 5\beta + \cos 5\gamma)$$

$$A_7 = \frac{2E}{21\pi R}\,(2\cos 7\alpha + \cos 7\beta + \cos 7\gamma)$$

Inspection of these equations shows that if the coefficients of the third, fifth, and seventh harmonics are set equal to zero, three simultaneous equations in these unknowns are formed. Solution of these equations should yield values of α, β, and γ which would eliminate these harmonics from the wave. A solution for these equations under which the coefficients of the third and fifth harmonics are reduced to zero was found when $\alpha = 18$ degrees, $\beta = 42$ degrees, and $\gamma = 78$ degrees, and these values were used in building the commutator.

The actual wave form of the motor coil current is shown in Fig. 7 directly below the idealized wave. It was recognized at the start that the inductance of the motor coils would not be negligible in any precise analysis of the system operation; however, this preliminary wave form was so satisfactory in operation that no changes from the original values of α, β, and γ have been required.

Production Design

With the basic operating principle of the inverter determined by the development, the production design of the precision low-frequency inverter for aircraft use involved four major considerations. These were:

1. Size and weight.
2. Radio noise suppression.
3. Operating life.
4. Vibration isolation.

Although the first of these, size and weight, enters into each part of the design, certain items control the basic size and weight of the unit. The main components of the inverter are the drive motor, brushes, the commutator, and the necessary gearing to accomplish the speed reduction from the 2,700 rpm of the d-c controlled speed motor to the 240 rpm of the commutator.

Table I. Inverter and Flowmeter Motor Operating Principle: See Fig. 4

Slip Ring Polarity			Coil Current			Resultant Vector	
1	2	3	1	2	3	Relative Magnitude	Direction
+	−	−	$+\frac{2E}{3R}$	$-\frac{1E}{3R}$	$-\frac{1E}{3R}$	$\frac{E}{R}$	0°
+	−	open	$+\frac{1E}{2R}$	$-\frac{1E}{2R}$	0	$0.866\frac{E}{R}$	30°
+	−	+	$+\frac{1E}{3R}$	$-\frac{2E}{3R}$	$+\frac{1E}{3R}$	$\frac{E}{R}$	60°
open	−	+	0	$-\frac{1E}{2R}$	$+\frac{1E}{2R}$	$0.866\frac{E}{R}$	90°
−	−	+	$-\frac{1E}{3R}$	$-\frac{1E}{3R}$	$+\frac{2E}{3R}$	$\frac{E}{R}$	120°
−	open	+	$-\frac{1E}{2R}$	0	$+\frac{1E}{2R}$	$0.866\frac{E}{R}$	150°
−	+	+	$-\frac{2E}{3R}$	$+\frac{1E}{3R}$	$+\frac{1E}{3R}$	$\frac{E}{R}$	180°
−	+	open	$-\frac{1E}{2R}$	$+\frac{1E}{2R}$	0	$0.866\frac{E}{R}$	210°
−	+	−	$-\frac{1E}{3R}$	$+\frac{2E}{3R}$	$-\frac{1E}{3R}$	$\frac{E}{R}$	240°
open	+	−	0	$+\frac{1E}{2R}$	$-\frac{1E}{2R}$	$0.866\frac{E}{R}$	270°
+	+	−	$+\frac{1E}{3R}$	$+\frac{1E}{3R}$	$-\frac{2E}{3R}$	$\frac{E}{R}$	300°
+	open	−	$+\frac{1E}{2R}$	0	$-\frac{1E}{2R}$	$0.866\frac{E}{R}$	330°
+	−	−	$+\frac{2E}{3R}$	$-\frac{1E}{3R}$	$-\frac{1E}{3R}$	$\frac{E}{R}$	360°

Coil current; $+$ = going toward center; $-$ = going away from center.

Since the size and weight of the motor are fairly well established by the availability of a motor with the required precision and since the size of the commutator and brushes is also established by the wave shape and output requirements, the problem is to obtain the best configuration in the least space with the least size and weight of supporting parts. As shown by Fig. 8, locating the brush and commutator arrangement back over the drive motor between two supporting plates gives a very compact assembly. The alignment of the commutator is obtained by the same blocks that hold the brush holders, and these blocks are of plastic material both for lightness and insulation of the brush holders.

Since the inverter is in a hermetically sealed case to provide dependable, long-lived operation independent of most ambient conditions, the case material must be of a material that will allow soldering. Normally this material is either brass or plated steel of sufficient thickness to withstand handling, as well as to provide sufficient attenuation for radio noise suppression. Even with a small case, either material of suitable thickness adds considerably to the weight of the unit. For this design, a copper-bonded aluminum, commercially known

as Alcuplate, consisting of 20% of thickness of copper bonded to aluminum on the outside of the case, is used for the case material. The use of this material provides the case with sufficient strength and permits soldering, yet is only approximately half the weight of brass or steel.

Radio Noise Suppression

Although the commutator spacing was designed to give a wave shape with the least amount of low-frequency harmonics, the inverter is still primarily a rotary
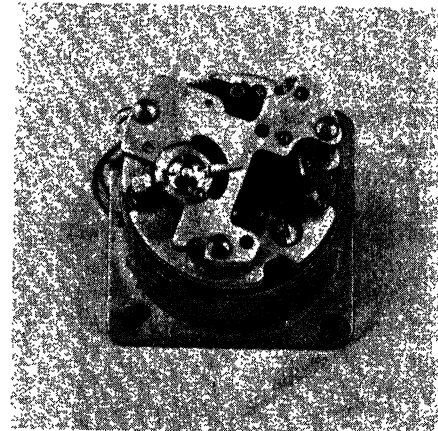
Fig. 5. Inverter driving motor
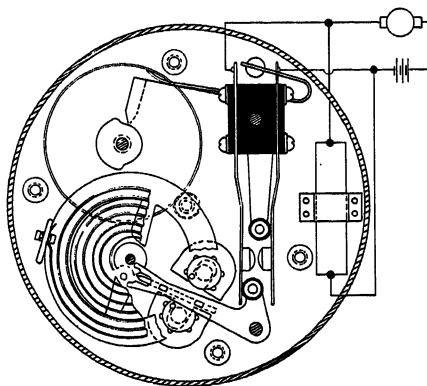
Fig. 6. Inverter driving motor details

Fig. 7. Flowmeter driving motor coil current

Table II. Output Frequency Tolerances

Test	Tolerance in Per Cent
Room temperature, 25 C............+0.15	
Full load...........................+0.15	
Voltage error.....................+0.05*	
Position error.....................+0.05*	
High temperature, +70 C..........+0.30	
Low temperature, −55 C............+0.30	
Vibration resonance................+0.15*	

* Change in frequency from room temperature output test.

switch driving an inductive load, and no actual commutation can be achieved. The resulting transients and arcing at the brushes from the switching create a very high noise level. Since these transients occur only every 1/48 second, there is a large difference between the peak and quasipeak values of radio noise, and a large amount of filtering is necessary to reduce the peak values below the limits specified in *MIL-I-6181B*.

Complete enclosure of the unit helps to contain the radiated noise, but filtering must be added to bring the conducted noise below specified limits and to prevent radiation from these lines. This filtering adds to the problem of size and weight of the unit. If conventional self-contained π filters in each line, external to the sealed inverter section, were used to accomplish the desired suppression, their size and weight would be equal to or greater than that of the inverter itself.

For this production design, the filtering was reduced to the basic elements of the conventional filters, and these elements were located within the sealed inverter case where lead lengths could be kept to a minimum and the components themselves placed at the most effective points. Thus, the toroidal inductors, Fig. 8, are located directly in the line between the noise source (brushes) and the feed-through capacitors, Fig. 8, which in turn feed directly into the output connector. Additional capacitance to minimize low-frequency noise was connected in parallel with the feed-through capacitors and was so arranged that the lead lengths were at a minimum. With such an arrangement, it was possible to use L sections in each line instead of π filters and still obtain the desired suppression. This means reduced filtering to one-sixth of the total size and weight of the unit. This accomplishment was aided considerably by some of the latest advances in electrical components, particularly capacitors.
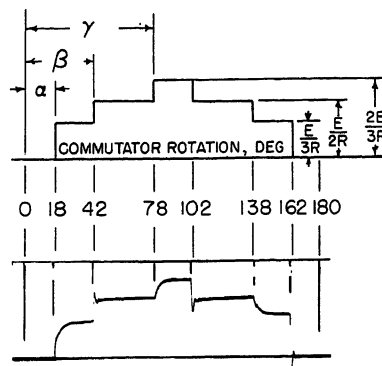
Operating Life

The problem of long operating life centered mainly about the brushes used in the inverter, not only those brushes used for switching but also those used in the drive motor. Selecting the proper brush for any application is a rather long and time-consuming project since tests cannot readily be accelerated, and this application is no exception. Very little work has been published on commutator speeds of 240 rpm by industry in general. The arcing of the switching, although reduced to some extent by the filtering, tends to shorten the brush life.

Since the inverter must operate under extremes of environmental conditions, it has been hermetically sealed to give a controlled atmosphere for brush operation. This procedure not only eliminates many variables in the selection of a brush but also gives a more constant condition for production units with resultant more reliable operation. Many tests were run on various kinds and types of brushes and one was finally found that filled all requirements, including a very low wear rate. This is an impregnated silver graphite brush and it is run in a dry helium atmosphere. Tests for over 2,500 hours of operation have been run

on these brushes. A similar program was conducted on the motor to insure its satisfactory operating life, as well as its starting and timing abilities after considerable hours of operation.

Vibration Isolation

Because the motor is dependent on the period of the balance wheel for proper timing, its timing can be affected by certain frequencies of vibration in certain planes. For this reason, the inverter must be vibration isolated to reduce the amplitude of these vibrations at the critical frequencies. A conventional amplifier type shock mount with a resonant frequency of 9 cycles and an efficiency of better than 65% over 26 cycles was used to reduce the amplitude of these vibrations. Under these conditions, timing errors do not exceed ±0.15% at the resonant frequencies of the motor.

Conclusion

The over-all result of this development and design program is a precision low-frequency inverter for an aircraft application with about a 40-watt capacity at 17 volts (average) and 4 cps which meets the frequency limits shown in Table II under the variables listed. The size of the unit is $3\frac{1}{2}$ by $4\frac{3}{4}$ by $4\frac{3}{8}$ and the weight is only $2\frac{1}{4}$ pounds. A completed power supply is shown in Fig. 9.

Fig. 8. Inverter construction details
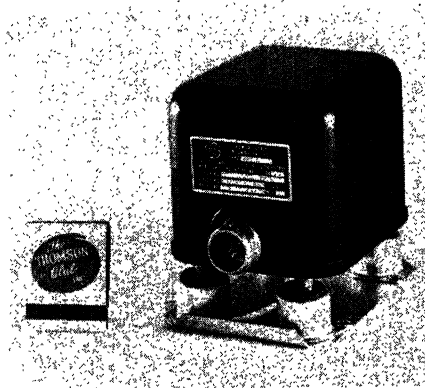
Fig. 9. Complete inverter

The use of this inverter in the mass rate flowmeter system for which it was designed virtually eliminates any error in the system due to transmitter impeller speed variation. In addition, transmitter weight and power requirements have been kept to a minimum.

References

1. INTERFERENCE LIMITS, TESTS, AND DESIGN REQUIREMENTS, AIRCRAFT ELECTRICAL AND ELECTRONIC EQUIPMENTS. Military Specification *MIL-I-6181B*, Washington, D. C., May 29, 1953.

2. MOMENTUM PRINCIPLE MEASURES MASS RATE OF FLOW, V. A. Orlando, F. B. Jennings. *Trans-actions*, American Society of Mechanical Engineers, New York, N. Y., Aug. 1954, vol. 76, pp. 961–65.

3. FUEL MASS FLOWMETER MEETS NEEDS OF AIRCRAFT OPERATION, H. T. Wrobel, R. F. Buckley. *Aviation Age*, New York, N. Y., Dec. 1953, vol. 20, p. 136.

4. PRECISION TIMER FOR D-C. *Electrical Manufacturing*, New York, N. Y., July 1949.

Correlation of the Thermodynamic and Electrical Characteristics of Blast-Cooled Generators

DANIEL FRIEDMAN
ASSOCIATE MEMBER AIEE

IN A recent paper a historical study was made of the procedures being used to correlate the thermodynamic and electrical characteristics of blast-cooled generators.[1] This paper will discuss the correlation procedures being used at the Naval Research Laboratory (NRL), explain why they were chosen and, where possible, present supporting experimental results.

For discussion purposes, the rating of blast-cooled generators will be divided into three phases: pressure-drop characteristics, heat transfer characteristics, and loss characteristics.

Nomenclature

a = proportionality constant
F = speed effect on generator pressure drop, inches of water
I = current, amperes
k = thermal conductivity of the cooling air
L = loss, watts
M = cooling air flow, pounds per minute
N = generator speed, revolutions per minute (rpm)
Nu = Nusselt number
P = absolute pressure, centimeters (cm) mercury

Δp = generator pressure drop, inches of water
Q = heat removed by cooling air, watts
Rn = Reynold number
T = absolute temperature, degrees Kelvin (K)
$\bar{T}$ = mean absolute temperature, K
α = exponent
γ = ratio of specific heats, for air $\gamma = 1.4$
ρ = air density, pounds per cubic foot[3]
μ = absolute viscosity

SUBSCRIPTS

a = average
b = brush
d = standstill
e = effective
i = inlet
o = out
R = rotor
t = total
s = stator
SL = reference sea level conditions
w = surface or wall

Pressure-Drop Characteristics

STANDSTILL CORRELATION

In an earlier NRL report, an equation was developed which made it possible to predict the effect of air weight flow, pressure altitude, load, and inlet air temperature on the generator pressure drop.[2] This equation is

$$\Delta p = a_1 M^2 / \rho \tag{1}$$

The large change in absolute pressure and temperature which the air experiences as it passes through the generator raises the question as to what correlation method should be used to calculate ρ.[3] To determine experimentally which correlation procedure best accounts for the effect of the cooling-air absolute pressure

on the generator pressure-drop characteristics, tests were conducted on a General Electric (GE) 15-kva generator (*2CM212A2*).[4]

Fig. 1 shows the standstill pressure-drop characteristics taken at sea level and at a pressure altitude of 50,000 feet with the use of two methods of calculating ρ.

Method A

$$\rho_a = a_2 / \left[\frac{T_i}{P_i} + \frac{T_o}{P_o} \right] \tag{2}$$

Method B

$$\rho_e = a_2 \frac{P_o}{T_i + T_o} \tag{3}$$

This use of outlet pressure is based on the results of equation 2, reference 5. At sea level $\rho_e = \rho_a$ since the drop in absolute pressure in the generator is negligible. Fig. 1 shows that method B gives the best correlation. Within the altitude and flow range covered, for the same pressure drop, the M^2/ρ_e for sea level and 50,000 feet (method B) agree within 4 per cent (%). This means that the predicted air flows would be within 2%.

To determine the effect of inlet air temperature, additional sea level tests were run at an inlet air temperature of −65 degrees centigrade (C). Using either correlation method, the low-temperature results were within the accuracy limits obtained in the foregoing. Additional tests, also shown in Fig. 2, show that either method successfully predicts the effect of load on the generator pressure-drop characteristics.

To improve correlation and reproducibility, almost all of the GE *2CM212A2* measurements were made with the generator wrapped in sufficient paper tape to make it completely airtight.

SPEED EFFECT

In addition to presenting a flow restriction, many generators are so constructed that pressure sources exist within the machine when the rotor is turning. This phenomenon is commonly known as the "fan effect." This fan effect may be the result of a fan built into the machine to provide cooling with no external pres-

Fig. 1. Effect of pressure altitude and inlet air temperature on generator pressure-drop characteristics: GE 2CM212A2, zero rpm, taped generator, axial inlet

pressure drop versus M^2/ρ curves. This may be due to a flow redistribution phenomenon within the generator or it may be due to an air flow effect on the fan characteristics. Since the two effects could not be separated, there was no way of determining what was happening inside the machine. Although the heat transfer correlations showed no flow redistribution effects, the results were not accurate enough to be conclusive.

Fig. 4 shows the results of tests made at two speeds and a number of pressure altitudes. To correlate the results shown in Figs. 3 and 4, the difference in pressure drop resulting from the speed effects within the machine were plotted as a function of speed and inlet air density. Fig. 5 shows the results obtained, which can be represented by the equation

$$F = -\left[1.0 + \frac{\Delta p_d}{32}\right]\rho_t^{0.6}\frac{\left(\dfrac{N}{1,000}\right)^{1.5}}{1.75} \qquad (5)$$

It should be pointed out that the results are valid only for the configuration investigated. For further details, see reference 4.

Where a high degree of accuracy is required, the effect of air flow, density, and speed on the speed effect should be determined experimentally for the specific configuration under consideration. For a rough approximation, however, equation 4 should be adequate.

sure drop, or it may be the result of characteristics inherent in the particular design of the machine.

For some machines, changing the speed of the generator may change the air flow resistance of the machine. This means a change in the value of a used in equation 1. Since the test results of almost all investigators can not separate the two effects, the "speed effect" is the algebraic sum of the two effects.

The measured pressure drop across the generator is equal to the actual internal pressure drop plus or minus the speed effect. It should be remembered that it is the actual internal pressure drop, not the measured over-all pressure drop, which should be used in equation 1.

Most investigators measure the speed effect at a sea level reference condition and calculate the effect for other conditions by application of the general fan laws.[1,3] In this case

$$F = F_{SL}\left(\frac{N}{N_{SL}}\right)^2\frac{\rho}{\rho_{SL}} \qquad (4)$$

For some types of generators, however, this assumption has been found to be invalid. Reference 6 indicates that for some generators the speed effect is a function of air flow and in some cases does not follow the speed correction shown in equation 4.

Similar results have been obtained at NRL using the GE 2CM212A2.[4] For these tests, a large fan was attached to the rotor shaft in order to provide a speed effect which was large enough to be

measured with reasonable accuracy. Fig. 3 shows the results of tests made at sea level with varying generator speed. Both types of speed effects seem to be operating here. First, all the lines do not pass through the intercept. This means that the fan acts as a pressure source. Second, there is a change in slope of the

Fig. 2. Effect of generator speed and load on its pressure-drop characteristics: GE 2CM212A2, 50,000-foot pressure altitude, no fan, taped generator, axial inlet
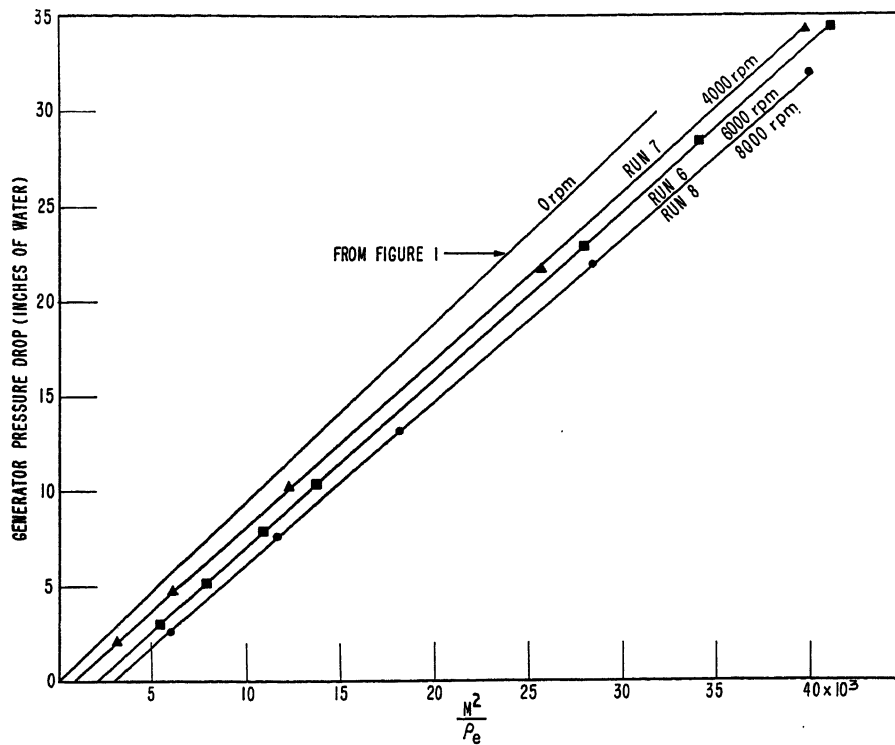
Fig. 3. Effect of generator speed on its pressure-drop characteristics: GE 2CM212A2, sea level, with fan, taped generator, axial inlet

COMPARISON OF CORRELATION METHODS

Equation 1 is basically the same as the equation which has been used by a number of investigators, namely[1,3]

$$M = a_2 \sqrt{\rho(\Delta p)} \qquad (6)$$

The relative merits of the two correlation methods can best be seen when it is attempted to determine the pressure-drop equations from experimental data. For identification purposes, the presentation method represented by equation 1 will be called method C; and the method represented by equation 6, method D. Method C requires a plot of pressure drop as a function of M^2/ρ and method D requires a plot of mass flow as a function of $\rho\Delta p$. If rectangular co-ordinates are used, the accuracy of method D is limited by the fact that the altitude curves are confined to the lower left corner of the plot. Since the curve or curves obtained are not straight lines, curve fairing and extrapolation are difficult. Although some of the resulting accuracy limitations can be eliminated by the use of log-log co-ordinates, the use of log co-ordinates makes it impossible to consider the negative pressure drops which may result from generator fan action. The use of method C also makes it easier to determine the speed effect, since this effect can be determined directly from the required figures. The use of method D, however, requires that the density term be taken into account.

COMPRESSIBILITY EFFECT

The correlation equations discussed so far are based on incompressible flow theory. At large pressure drops, the compressibility effects must be considered.[1] Reference 1 points out that existing information does not cover a $\Delta p/P$ range large enough for any conclusions to be derived.

GRAPHICAL PRESENTATION

For convenient use, the rating equations can be presented in graphical form. This makes it possible to predict the rating of the generator, and for some pres-

Fig. 4. Effect of pressure altitude on generator pressure-drop characteristics: GE 2CM212A2, 6,000 rpm, with fan, taped generator, axial inlet

entation methods, the air flow and heat rejection, without having to make any calculations.[1] Considering the complexity of the problem, this will result in the saving of considerable time and effort when designing generator installations.

In the 3-chart presentation method, see Fig. 6, the air flow is plotted as a function of pressure drop and generator outlet static pressure. Equation 1, however, indicates that the pressure-drop characteristics are also a function of air temperature. For most installations this problem can be solved by making the following simplifying assumptions:

1. The static air temperature as a function of altitude is obtained from the tabulation of standard hot atmosphere or similar temperature-altitude relationship.[7]

2. The ram temperature rise, for a given set of conditions, is calculated assuming that the pressure drop across the generator is one-half of the free stream dynamic pressure. In equation form

$$P_t = P + 2\Delta p \qquad (7)$$

From the Appendix of reference 8

$$\frac{T}{T_t} = \left(\frac{P}{P_t}\right)^{\frac{\gamma-1}{\gamma}} \qquad (8)$$

or for air

$$\frac{T}{T_t} = \left(\frac{P}{P_t}\right)^{0.286} \qquad (9)$$

In this case $T_i = T_t$.

3. $T_o - T_i = \dfrac{\text{loss at rated load in watts}}{7.6M} \qquad (10)$

4. The generator pressure-drop characteristics can be represented by equations 1 and 4 or their equivalent.

Where refrigeration, engine bleed-air, or similar devices are used, the inlet air temperature is not directly related to stag-

Fig. 5. "Speed effect" as a function of generator speed and air density: GE 2CM212A2, with fan, taped generator, axial inlet

nation air temperature. A method of presentation has been suggested by F. M. Potter which overcomes this difficulty.

In this case, the air flow is plotted as a function of pressure drop and an arbitrary reference density. This reference density is determined graphically from the inlet air temperature and outlet static pressure. No co-ordinates for this reference density need be shown. The pressure-drop lines should be calculated with the use of the average temperature for the range covered by the chart and the calculation procedures discussed in the foregoing. With this presentation method, a variation in inlet temperature from −55 to +80 C will introduce an error of about 2%. Fig. 7 is a plot of the characteristics of the GE 2CM212A2 generator using this presentation method.[4]

The presentation method to be used for a particular application depends on the type of configuration involved and the accuracy desired. The second presentation method is more accurate than the first, but it involves a great deal more calculation. The additional work required by the last presentation method is justified only when inlet temperatures are nowhere near normal (variation greater than 25 C) ram temperatures.

Heat-Transfer Characteristics

COEFFICIENT OF HEAT TRANSFER

Forced convection heat transfer is frequently correlated by the equation[1,3]

$$Q = aM^\alpha(T_e - T_i) \qquad (11)$$

For T_e, a number of investigators have used equivalent or measured surface tem-

peratures.[3] Since it is the hot spot which limits component life, it seems logical to use the hot-spot temperature for T_e. Otherwise it will be necessary to obtain a correlation between hot-spot temperature and the reference value used for T_e. At NRL, the hot-spot temperatures are used in equation 11.

The next question is how hot can T_e be allowed to go? For most components, this value is a function of the life desired, mechanical stress, assembly methods, and other mechanical considerations. Since different materials and construction

techniques are used by different manufacturers for the same type of equipment, the best source of this type of information is the equipment manufacturer, who, it is expected, will obtain this information from their material suppliers, from additional life tests, and from measurements made on machines which have proved satisfactory in actual use.

The constant α must be determined experimentally for each generator. For the same machine, the value of α is a function of the choice of correlation temperatures. For example, the use of hot-spot temperature will usually give a smaller value of α than would be obtained using surface temperatures.

If changes in air temperature and air flow shift the limiting temperature from one component to another, it may be necessary to develop equations for each component. For example, the rotor windings may be limiting at one condition, the stator windings at another, and the bearings at a third condition. It will be necessary, when determining the rating for a specific set of conditions, to check the temperature of all critical components to determine which one is limiting.

ROTOR SPEED AND ALTITUDE

The air flow M to be used in equation 11 is the flow past the heat-transfer surface under consideration. It has generally been assumed that the flow rate past every component under investigation is a direct function of the total air flow. In

Fig. 6. Generator altitude rating chart: GE 2CM212A2, 6,000 rpm, 0.75 power factor, no fan, axial inlet, untaped generator

Fig. 7 (left). Graphical presentation of generator pressure-drop characteristics: GE 2CM-212A2, with fan, axial inlet, untaped generator

Fig. 8 (below). Stator heat-transfer correlation based on total loss: GE 2CM212A2, axial inlet, taped generator

To study the phenomena involved, it will be best to start with the simplest case, a cylinder rotating in free air. For this configuration[9]

$$Nu = 0.1 \, Rn^{2/3} \qquad (12)$$

In terms of the variables used in this report, equation 12 reduces to

$$\frac{1}{k}\left(\frac{L_R}{T_w - T_i}\right) = a\left(\frac{N\rho}{\mu}\right) \qquad (13)$$

(The physical dimensions are included in the constant.)

Within the range of -55 C to $+80$ C and sea level to a 65,000-foot pressure altitude, this reduces to

$$\frac{T_w - T_i}{L_R} = a(NP)^{-0.67}(T)^{+0.32} \qquad (14)$$

Luke[10] experimentally studied the heat transfer from a rotating cylinder with a circular housing around it. With no air flow, his experimental results can be expressed by

$$\frac{T_w - T_i}{L_R} = a_N{}^{-0.62} \qquad (15)$$

With air flow through the gap between the rotor and the outer cylinder, the exponent varied from -0.62 to -0.2, depending on the ratio of rotor surface speed to the air through velocity. For all speeds, the through air flow was the same.

Equations 14 and 15 apply to surface temperature measurements. The inclusion of insulation resistance between the wire and the outside air, and the use of hot-spot temperatures rather than average surface temperatures may change the coefficients of temperature, speed, and pressure.

For actual generators, the investigation of these effects is complicated by the fact that a large number of other factors are affected by changes in speed and altitude. For example, increasing the generator speed increases generator friction losses. As a result, the air reaching the rotor may be hotter.

Tests of the GE 2CM212A2 generator produced test results which showed pressure and temperature effects consistent with equation 14, but the predicted speed and air flow effects were not observed.[3,4] For the rotor

$$\frac{T_R - T_i}{L_R} = \frac{1.805}{(P_a M)^{0.4}} \qquad (16)$$

This equation indicates that the rotor heat-transfer coefficient at 50,000-foot pressure altitude is about one-half sea level value. It would seem, until further experimental evidence has been obtained, that rating tests must be taken at altitude as well as sea level.

many cases, this assumption is not valid. For example, assuming a negligible air flow through a generator, there would be an air flow over the rotor surfaces due to the relative velocity of the rotor surface to the air in the generator. Under operating conditions, the air flow over the rotor surfaces is the resultant of the axial and peripheral air flows. The peripheral air flow is a function of rotor rpm, absolute air pressure, and air temperature. When the axial velocity is of the same order of magnitude as the peripheral velocity, the peripheral air flow should produce a speed, altitude, and temperature effect on the heat-transfer coefficient. This effect should be greatest at reduced air flows.

Fig. 9. Stator heat-transfer correlation based on estimated stator loss: GE 2CM212A2, axial inlet, taped generator

The rotor temperatures used in equation 16 are average temperatures calculated from resistance measurements. These measurements were made while the generator was running with the use of the bridge circuit discussed in reference 11.

It should be pointed out that the exponents used in equation 14 will vary from one generator to another. A generator designed to have high velocities for a given mass flow would tend to show less altitude effect than one designed to have low air velocities. As a general rule, for machines with the same nominal rating, the one with the highest pressure drop for a given flow will have the smallest altitude effect.

Loss Correlation

Previous investigators have used the generator total loss as the Q in equation 11. For some components, a better correlation can be obtained if Q is taken to be the loss within the component. For example, the *2CM212A2* tests indicate that rotor losses should be used to correlate rotor temperature.[4]

An interesting example of how important it is to choose the correct loss for correlation purposes was obtained when analyzing the stator temperature measurements taken on the *2CM212A2*. Fig. 8 shows the heat transfer correlation of stator measurements using total losses as Q. There appears to be a definite altitude effect. When the estimated stator loss was used instead of the total loss, no altitude effect could be observed; see

Fig. 9. Both loss values seemed to take into account the effect of speed and power factor.

In attempting to correlate brush heat-transfer phenomena, no temperature rise per unit loss correlation seemed satisfactory. In addition, factors other than speed, brush current, air flow, and air temperature seemed to affect brush temperature, which, for example, is known to be affected by humidity and by previous operating history.

The temperature rise for each test was plotted as a function of air flow on log-

log paper. The slopes varied from -0.2 to -0.5 with most of the data at the smaller slopes. Rather than present the large mass of data, which would be more confusing than enlightening, a cross plot of these results is shown in Fig. 10. The two parallel lines seem to bracket the results. For correlation purposes, the upper line is used, whose equation is

$$T_b - T_i = 10.9(I_R)^{2/3} \tag{17}$$

As an approximate average figure, an exponent of $-1/4$ will be used for air flow. The equation for brush rise becomes

$$T_b - T_i = 19.6(I_R)^{2/3}(M)^{-1/4} \tag{18}$$

Flow Redistribution

By changing the flow distribution within the machine, internal fan action may change the effective heat transfer coefficient within the machine. For the *2CM212A2* generator, however, attaching the fan to the inlet end of the drive shaft produced no effect on any of the heat transfer coefficients within the machine. For other generators, however, this effect may not be negligible.

Experimental Considerations

Using the correlation methods outlined, it is possible to make a complete thermal evaluation of a generator with a limited number of tests. In addition, these correlation methods do not require that temperature stabilization be achieved at conditions where one of the components is operating within a few degrees of its maximum allowable temperature. This means that the required data can be taken at lower temperature levels with
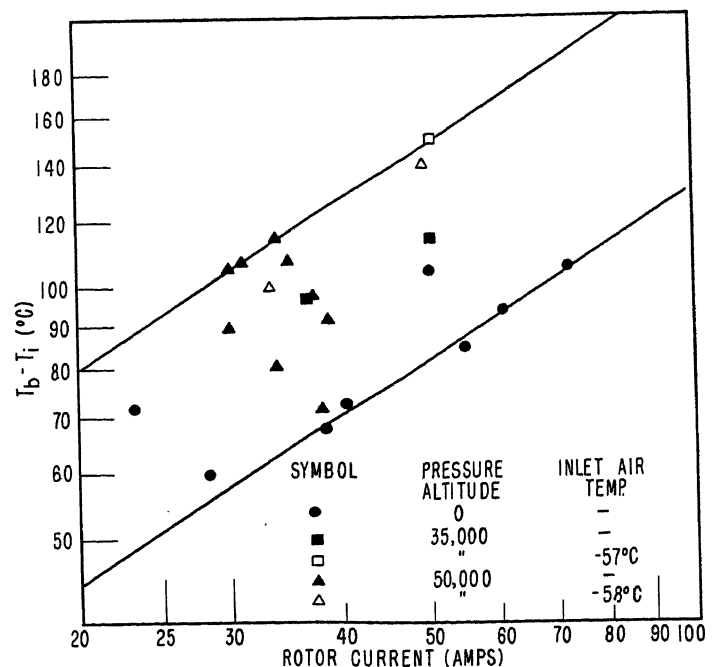
Fig. 10. Brush temperature rise as a function of rotor current: GE 2CM-212A2, axial inlet, taped generator, air flow of 10 pounds per minute

less expenditure in time and money. By overloading and underloading the generator for short periods, the average test time can be further reduced.

At the present state of the art, each organization doing experimental thermal evaluation has developed its own instrumentation techniques. As test results become available, sufficient information will be obtained with which to judge the relative merits of the different techniques being used. This comparison is handicapped, however, by the fact that very few reports adequately discuss the techniques and instrumentation used.

OTHER HEAT FLOW PATHS

It should be remembered that heat can be added to or removed from the generator through other heat flow paths. Free convection, radiation, and conduction through the mounting pad must be considered. The magnitude and direction of this heat flow depends on the environmental conditions of the particular installation. Almost all rating studies have assumed that all the losses were removed by the cooling air. This assumption was made since it was felt that the ratings which were arrived at were on the safe side. In addition, the thermally insulated condition serves as a convenient reference point from which rerating equations for other environmental conditions could be developed.

In many experimental thermal evaluations, however, test conditions were not set up with these factors in mind. For example, tests have been run with the generator mounted directly to a cold pad. In this case, a rating chart based on these data would not be on the conservative side. Rating tests are run at NRL with the outside of the generator thermally lagged and the generator insulated from the drive stand. Tests were run with the generator surface insulation removed to determine the effect of these heat-flow paths on the temperatures within the generator. For the 2CM212A2 generator, even with an external air velocity of about 25 mph, no change in internal temperatures could be detected.

In this installation, the drive-stand insulation could not be removed. Tests were run on an instrumented 2CM73C4 generator to determine the effect of pad heat conduction.[12] The results obtained show that the temperatures within the generator, particularly the bearing temperature, are affected by pad temperature.

Robinson, Barnum, and Buxton have also considered these factors in their thermal study.[13] The reports of many other

investigators, however, do not discuss the installation in sufficient detail to be able to tell what precautions were taken to take these factors into consideration.

TRANSIENT CONDITIONS

For short-time applications, the transient response to a step change in air temperature and/or air flow is important. For the chamber used, it was impossible to change the air temperature rapidly, and the change in air flow used was far from a step function. Preliminary data indicate that the time constant for rotor and stator was about $6\frac{1}{2}$ minutes, and about 2 minutes for the brushes. Complete equilibrium was achieved in about 1/2 hour. These data were obtained during the regular testing procedure with a minimum amount of added test time.

Loss Characteristics

To be able to predict the rating of a generator, it is necessary to know what the loss characteristics of the generator are. This can be calculated from the design of the machine or determined experimentally.

When component temperatures are best correlated against the losses within the component, a method for determining these losses must be established. For example, for a-c machines the rotor losses are not directly related to the overall losses. Standard test and calculation procedures are available that can be used to estimate the loss distribution within a-c and d-c generators. For a detailed discussion of how this was done experimentally for one specific generator, see reference 4.

GRAPHICAL PRESENTATION

As was pointed out earlier, the pressure-drop, heat-transfer, and loss characteristics can be presented in the form of a rating chart which can be used to predict the rating of a machine for any set of environmental and operating conditions.

In the 3-chart presentation method shown in Fig. 6, the curves in the upper right section are plotted from the heat-transfer correlation equations. The ordinate of these curves is the air flow through the generator to the same scale as is used for the upper left section of the figure. The abscissa is the generator or component loss used for correlation purposes. The lower right section is a plot of generator rating as a function of the reference loss used in the upper right section.

Where speed effects are important (gen-

erator speed variation greater than 5%) separate charts should be prepared for minimum and maximum rated speed. This would take into account the change in loss distribution as well as any change in heat-transfer coefficient with speed.

For some machines, the effect of power factor can be taken into account by plotting additional curves in the lower right part of the rating chart. Where there is an appreciable change in loss distribution within the machine, it may be necessary to have a separate chart for each power factor.

An altitude effect on the heat-transfer characteristics can be taken care of by having a series of lines in the lower right section of the chart. Fig. 6 shows the result of such a presentation for the 2CM21-2A2. A brief description of how this figure was prepared will best explain how the presentation method works.

1. For each generator operating condition, the loss distribution is determined within the machine.[4]

2. For at least three mass flows within the range under consideration, the allowable air temperature was calculated for each component using the experimentally determined correlation equations.

3. Steps 1 and 2 were repeated for at least two other loads.

4. For the rotor correlation, calculations must be made for several outlet pressures.

5. For the highest pressure altitude considered, a plot of allowable inlet air temperature versus the alternator rotor copper losses was prepared for all critical components. A separate figure was prepared for each mass flow. These figures are a graphical determination of what components were limiting for a given set of conditions. The curves shown in the upper right section of Fig. 6 are a cross plot of these figures. Although no scale is shown, the abscissas of the two right plots are the log of the alternator rotor loss. The decision as to which altitude to use in the determination of the upper right plot is a somewhat arbitrary one. The maximum altitude condition was chosen since it had the least number of discontinuities.

6. In the lower right section of the figure, the curve representing generator rating as a function of the reference loss was plotted. This is the 50,000-foot line shown.

7. For the other altitudes, the allowable inlet air temperatures were plotted as a function of the allowable output. Projecting straight down from the intersection point of the air weight flow and air inlet temperature, a point was located in the lower right plot of air temperature versus output. For each altitude a curve was drawn which represented the upper right boundary of the points so obtained. The results will therefore be on the conservative side.

8. Since for some applications the total loss is required, a plot of total loss versus output for two altitudes was also included.

Conclusions

GENERAL

1. The process of rating blast-cooled generators can be divided into three phases: pressure-drop characteristics, the heat-transfer characteristics within the machine, and the relation between generator load and generator losses.

2. Many investigators in the field are not supplying sufficient information concerning test and instrumentation procedures to enable others to evaluate properly their results.

3. Except for some specialized requirements, generator rating charts should be prepared in the form shown in Fig. 6. In cases where speed is important (speed variation greater than 5%), two or more charts should be prepared for the same generator.

PRESSURE-DROP CHARACTERISTICS

1. For a generator without a speed effect, equation 1 should be used to correlate the pressure-drop characteristics. The density term in the equation should be calculated from average air temperature and outlet static pressure. This equation takes into account the effect of air temperature, air flow, pressure altitude, and generator load on the pressure-drop characteristics of the generator.

Where there is a speed effect, the change in pressure-drop characteristics should be experimentally determined for a number of speeds, altitudes, and flow rates. If this is not possible, equation 4 can be used to estimate the effect of air density and generator speed. The pressure drop to be used in equation 1 is the pressure drop which would exist if the generator was not turning, not the measured over-all pressure drop.

2. Sufficient data are not available with which to predict the pressure-drop characteristics of generators in the compressible flow range.

HEAT-TRANSFER CHARACTERISTICS

1. The limited experimental data available seem to indicate that for some generators, general speed and pressure altitude affect the heat-transfer coefficient. When developing rating equations for a specific generator, the effect of these variables should be experimentally evaluated. If these effects can be neglected, equation 11 should be used to correlate the heat-transfer characteristics of the generator.

2. Until experimental evidence is obtained to justify another correlation method, hot-spot temperature should be used in heat-transfer equations.

3. For some machines, separate rating equations may have to be developed for several components within the machine (bearings, rotor, stator, etc.).

4. The transient response to a step function change in cooling can be determined during steady-state evaluations with little added complexity and test time required.

5. During experimental evaluations, the effect of other heat flow paths on the generator temperature distribution should be considered.

6. For some machines, component temperatures are best correlated against the losses within the component, rather than over-all losses.

7. The correlation methods discussed in this report do not require that temperature stabilization be achieved at conditions where one of the components is operating within a few degrees of its maximum allowable temperature. This means that the required data can be taken at lower temperature levels using less time and money.

References

1. TEN YEARS OF PROGRESS IN PREDICTING THE AERODYNAMIC, THERMODYNAMIC, AND OUTPUT CHARACTERISTICS OF BLAST-COOLED AIRCRAFT GENERATORS, D. H. Scott. *AIEE Transactions*, vol. 73, pt. II, 1954 (Jan. 1955 section), pp. 455–61.

2. MEASUREMENT OF AIR FLOW THROUGH AN AIRCRAFT GENERATOR DURING FLIGHT, PART II—ANALYSIS OF FLIGHT TEST DATA FOR F7F-3 AIRCRAFT, J. M. Marzolf, D. Friedman. *NRL Report 4048*, Naval Research Laboratory, Washington, D. C., Sept. 1952.

3. CORRELATION OF THE THERMODYNAMIC AND ELECTRICAL CHARACTERISTICS OF BLAST-COOLED AIRCRAFT GENERATORS, PART I—BACKGROUND INFORMATION AND PRESENT STATE OF THE ART, D. Friedman, D. H. Scott. *NRL Report 4609, Ibid.*, 1955.

4. CORRELATION OF THE THERMODYNAMIC AND ELECTRICAL CHARACTERISTICS OF BLAST-COOLED AIRCRAFT GENERATOR, PART II—EXPERIMENTAL EVALUATION OF A GENERAL ELECTRIC 15-KVA NARROW-SPEED RANGE GENERATOR (MODEL 2CM212A2), D. Friedman. *NRL Report 4610, Ibid.*, 1955.

5. CORRELATION OF RATING DATA FOR ROTATING ELECTRIC MACHINERY, Erwin O. A. Naumann. *AIEE Transactions*, vol. 73, pt. II, 1954 (Jan. 1955 section), pp. 443–52.

6. HEAT-TRANSFER PROCESS WITHIN ROTATING ELECTRICAL EQUIPMENT FOR AIRCRAFT, J. R. Barnum, O. E. Buxton, W. Robinson. Ohio State University Research Foundation, Columbus, Ohio, AF Project 445, *Report No. 3*, July 1952.

7. ATMOSPHERIC PROPERTIES—EXTREME COLD AND HOT; STANDARD FOR AERONAUTICAL DESIGN. *Air Force-Navy Aeronautical Bulletin No. 421*, Washington, D. C., Sept. 1953.

8. COMPRESSIBLE FLOW TABLES FOR AIR, Marie A. Burcher. *NACA Technical Note 1592*, National Advisory Committee for Aeronautics, Washington, D. C., Aug. 1948.

9. CONVECTION FROM AN ISOLATED HEATED HORIZONTAL CYLINDER ROTATING ABOUT ITS AXIS, J. T. Anderson, O. A. Saunders. *Proceedings, Royal Society*, London, England, series A, vol. 217, no. 1131, pp. 555–62, May 21, 1953.

10. THE COOLING OF ELECTRIC MACHINES, George E. Luke. *AIEE Transactions*, vol. 42, June 1923, pp. 636–52.

11. MEASUREMENT OF THE AVERAGE TEMPERATURE OF AIRCRAFT A-C GENERATOR WINDINGS UNDER ACTUAL OPERATING CONDITIONS, D. Friedman. *NRL Memorandum Report 294*, Naval Research Laboratory, Washington, D. C., Apr. 1954.

12. EFFECTS OF MOUNTING-PAD TEMPERATURE ON THE RATING OF BLAST-COOLED AIRCRAFT D-C GENERATORS, J. M. Marzolf. *NRL Memorandum Report 428*, Naval Research Laboratory, Washington, D. C., Feb. 1955.

13. THERMAL STUDY OF A BLAST-COOLED DIRECT-CURRENT AIRCRAFT GENERATOR, W. Robinson, J. R. Barnum, O. E. Buxton. *WADC Technical Report 53–368*, Wright-Patterson Air Force Base, Wright Air Development Center, Ohio, Dec. 1953.

14. MEASUREMENT OF AIR FLOW THROUGH AN AIRCRAFT GENERATOR DURING FLIGHT, PART I—INSTALLATION AND CALIBRATION OF INSTRUMENTATION SYSTEM, J. M. Marzolf. *NRL Report 3961, Ibid.*, Apr. 1952.

Discussion

R. A. Yereance (Boeing Airplane Company, Seattle, Wash.): This is a very good paper and presents some points with which we are in complete agreement. This is particularly true of the statement following equation 16 that, "It would seem, until further experimental evidence has been obtained, that rating tests must be taken at altitude as well as sea level." We also agree that the value of some reports has been decreased because insufficient information was given as to test techniques.

We do not agree with all of Mr. Friedman's statements but believe this is due to the fact that our tests were run on 60-kva 400-cycle 3-phase constant-speed machines, much larger machines than those tested by Mr. Friedman, and also due to the fact that we were not free to make changes in the machines tested, as was Mr. Friedman.

In this paper it is stated that the $M^2/\rho e$ curves for sea level and 50,000 feet agree within 4% within the altitude and flow range covered. However, it is also stated in several places that there may be flow distribution changes through the machines under varying conditions. We believe that such flow changes occurred with changing altitude in the machines we tested in the range of normal cooling air flow, and we have not found good correlation between our sea level and altitude $M^2/\rho e$ curves, possibly for this reason. This would add further weight to the statement that, in the present state of the art of blast-cooled machines, tests must be run at altitude as well as at sea level.

In equation 11 it is suggested that hot-spot temperatures be used. The hottest components of the machines we have tested have been the rotors, and we have to date, found no practical method of determining the rotor hot spot while the machine is running, as we were not free to add additional instrumentation slip rings to the machine.

Preceding equation 7, it is stated that the ram temperature rise was calculated assuming that the pressure drop across the generator was one half of the free-stream dynamic pressure. It should be pointed out that the temperature rise is due to the pressure drop across the generator and ducting system, and not the generator alone. In our work we assumed that the temperature rise would be 85% of the theoretical temperature rise for 100% ram recovery and data obtained on our airplanes since then have seemed to indicate that this value may have been too low. For some installations assuming 100% of the theoretical temperature rise may not be overly pessimistic.

This report shows the results of good and extensive experimental work and analysis of the data obtained. Our remarks are intended merely to point out that the

work was done on a single machine and installation and that, as sufficient background in this field has not yet been gained, the results can not be applied to any other situation.

John Alger and J. T. Bateman (General Electric Company, Erie, Pa.): A significant step in the altitude rating of aircraft generators has been the adoption by the industry of a uniform format for rating charts. The more difficult step has yet to be taken: that is, to define the assumptions and compromises which may be made in the interest of simplifying the chart and minimizing the amount of calculation and costly testing that go into its preparation.

The author of this paper has certainly illuminated the way for that second step. Rather than presenting a single approach, he had compared the predictions of many possible assumptions with the results of thorough and exhaustive altitude testing on a production generator. Correlations are shown clearly enough to serve as a guide to machine improvement.

The graphical presentation of air flow by factoring in standard atmospheric temperatures and ram temperature rise is an excellent means of increasing the accuracy of the chart without additional complexity. There is some indication that pressure drop varies with the mass flow to a power slightly less than two in particular generator designs. The characteristic of a particular machine should be obtained by test rather than assuming that the machine behaves as a series of orifices.

By using the loss in the limiting component instead of over-all machine loss, the author has obtained a close correlation of data which leads to significant improvement in testing technique. Altitude tests may be run to provide a substantial (but not a specific) temperature rise and later corrected to the ultimate component temperature. This temperature rise per unit loss correlation can cut altitude testing time in half.

Daniel Friedman: The author wishes to thank Messrs. Yereance, Alger, and Bateman for their comments.

I agree with Messrs. Alger and Bateman's comments on pressure-drop characteristics. A great deal of experimental data have become available since the preparation of this paper. This information seems to indicate that

$$\Delta p = \alpha_1 (M^2/\rho)^b$$

where b is an exponent with a numerical value somewhere between 1.0 and 0.9.

There seems to be a need for further clarification of the intent of equations 7 and 8. Equation 7 is an estimate of the difference between the stagnation (total) pressure and the static pressure for the aircraft since it is the speed of the aircraft which determines the ram temperature rise. Equation 8 assumes that the air temperature rise is 100% of the theoretical ram temperature rise.

An Oil-Cooled A-C Generator for High-Speed High-Altitude Aircraft

H. J. BRAUN
MEMBER AIEE

W. J. SHILLING
ASSOCIATE MEMBER AIEE

Synopsis: A new approach to the problems of brush life and commutation, bearing life and lubrication, and high ambient temperature in aircraft generator design is presented. A test and development program for obtaining adequate data for the design and construction of an oil-cooled machine is outlined. The prototype machine and some construction details are shown.

IN THE aircraft industry, many novel methods are tried in an attempt to solve perennial problems. This paper presents a new approach to three of the ever-recurring problems of the aircraft generator designer.

Since the invention of the wheel, bearings and their lubrication have been a problem. For most surface equipment, fairly adequate solutions have been found.[1] In aircraft, however, with the extreme variation of temperature and moisture conditions, with the demands for light weight, and with the high-amplitude vibration conditions caused by a high engine horsepower to airplane weight ratio, bearings and their lubrication account for a large percentage of the unsatisfactory reports for aircraft generators.[2] Weight competition, the necessity for operation in all positions, and the necessity of keeping brushes free of oil have prevented the use of oil-lubricated bearings which, it is believed, would eliminate much of the bearing problem.

A second problem of the electrical designer on electric equipment has been the problem of brushes. Again, fairly acceptable solutions have been found for surface equipment. In aircraft, the design compromise necessary to take care of the variations in temperature, altitude, and humidity have been so great that in general only fair to poor operation is obtained under any one condition. A considerable amount of work has been expended to obtain these compromises, yet brushes and their life continue to be another source of unsatisfactory reports on aircraft generators.[3-8]

A third problem results from the ever-increasing speed of aircraft. This speed raises the temperature of air brought on board the aircraft to cool the equipment.[9-11] The power expended to get the air on board rises to astronomical values for small amounts of air, which at these higher temperatures tends to be much less effective. Fig. 1 shows the stagnation temperature rise of air for various speeds and the power expended to get a quantity of air on board. The quantity of air shown is approximately that allowed for a 60-kva machine by the new military class C specifications. Fig. 2 shows the trend in cooling coefficients, air rise, and copper loss with increasing temperature. Fig. 3 shows the life characteristics of that group of insulations designated as class H.[12] This curve represents the present state of the insulation art applicable to electrical machinery.

Discussion

In an attempt to solve these three problems, a development program was started to arrive at an aircraft generator based on new concepts. It was believed that these concepts would lead to a workable solution. The new concepts are listed as follows:

1. Oil-lubricated bearings would greatly minimize or reduce the bearing problem.

2. Elimination of brushes would eliminate the brush problem and greatly simplify the oil-lubrication problem.

3. Use of a different heat sump or a different method of cooling would take care of the temperature problem.

When it was decided to try to build an aircraft generator based upon different concepts than those used in the past, a
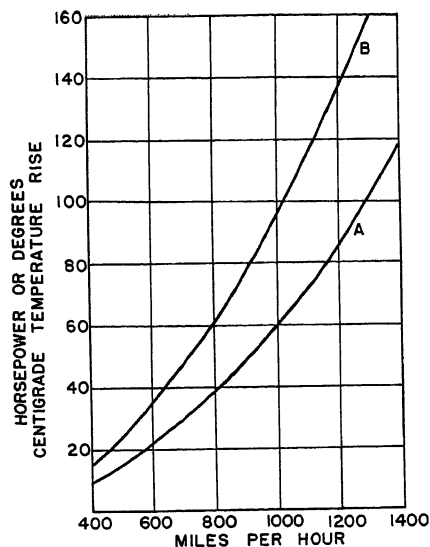
Fig. 1. Horsepower cost and stagnation air temperature rise versus flight speed

Curve A—Horsepower required to take 1 pound of air per second on board an aircraft

Curve B—Stagnation air temperature rise

Fig. 2. Temperature effects in air-cooled generator

Curve A—Heat transfer coefficient with constant head and altitude versus cooling air temperature

Curve B—Air temperature rise with constant head, altitude, and constant losses versus cooling air temperature

Curve C—Copper loss with constant current density as function of copper temperature

Fig. 3. Life of class H insulation

number of questions had to be answered. Some of the answers are of necessity arbitrary, others are based on study of the literature, and still others on tests. A list of questions follows. The answers will be discussed in the text following the list.

1. What rating should be experimented with first?

2. Should the machine and its system be independent of any battery or d-c system?

3. What coolant should be used?

4. How should the coolant be used?

5. What should the coolant flow rate be?

6. What should the design temperatures for the coolant be?

7. Can seals be relied upon in machine after machine?

8. What materials are compatible with the coolant at the design temperature?

9. What materials are compatible with the coolant vapor at the design temperature?

10. What type of machine should be built?

11. What excitation system should be used with the machine?

12. If rectifiers are used with the excitation system or the machine, what are the general requirements?

13. What type of rectifier should be used?

The rating of a machine which is to be built is arbitrary and based largely upon economics. Past economic experience of the authors' company indicated an excellent market potential, in general, for a-c aircraft generators rated at 40 kva. It was therefore decided to start develop-

ment on a 3-phase machine to be rated 40 kva, 120/208 volts, 400 cycles.

Other performance criteria were to be in line with military specification MIL-G-6099[13] when applicable. Later after appreciable testing had been completed and the type of machine had been decided, negotiations with a particular potential user altered the short-circuit capacity from the 300 per cent (%) required by MIL-G-6099 to 400%, and changed the specification of unbalanced voltage with unbalanced loading.[13]

Early in the development it was decided, again arbitrarily, that if possible, this machine or its system should fulfill the dreams of the a-c systems designer of a batteryless a-c aircraft power system. In the past, because of weight considerations, most relaying in a-c aircraft systems has been actuated by direct current. This direct current has been obtained from a transformer-rectifier system feeding a battery. The battery is heavy for the two functions it serves in an operating system. First, it furnishes excitation power to insure reliable starting of the a-c generator, and second, it furnishes relaying power during a fault. A small permanent-magnet generator with a rec-

tifier would serve the same two purposes and would weigh considerably less than a battery.
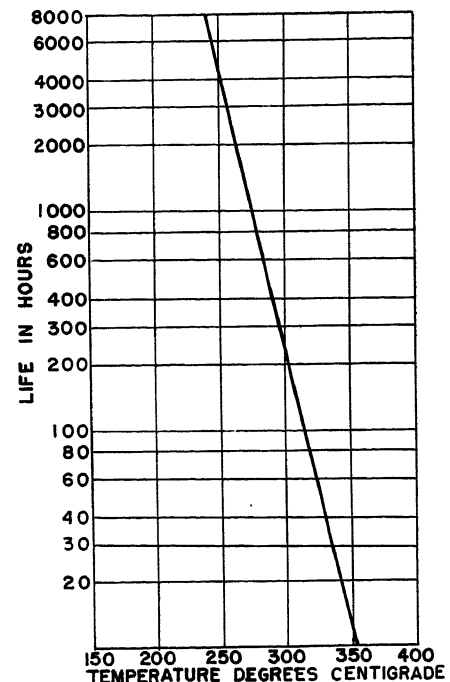
The problem of the coolant to be used is a complex mixture of engineering, psychology, and convenience.[14,15] A fairly complete survey of possible coolants for aircraft generators was made in unpublished work by Martin and Hambor. The authors' engineering associates surveyed and evaluated the field of coolants and decided that for aircraft generators on planes as built at present the two logical coolants other than air are fuel and oil. Fuel has many things to recommend it. It is on board modern aircraft in comparative abundance. It has good thermal conductivity and a low, relatively constant viscosity. It is already being used to cool the engine oil under most flight conditions. If oil were to be used, fuel would be the final sump under high-speed conditions, and fuel and air would be the sump under low-speed conditions. Oil has many things to recommend it, also. Aircraft a-c generators are used with some auxiliary drive or transmission to give constant output frequency. These drives require oil for their bearings, and the hydraulic type requires oil for their operation. These drives contain pumps for moving the oil. If oil were used, these same pumps could be expanded to move oil through the a-c generator. If oil is used as a coolant, the same oil can be used for the bearings. The aircraft engine uses oil to cool its bearings. Interchangers are available to cool this oil

Fig. 4. Test setup used for testing insulation materials in oil (the ruler is 18 inches in length)

for the engine. Many engines have already integrated into the oil system a constant-speed drive system. Little increase in the interchanger size would therefore be necessary to handle the a-c generator losses. Finally, a psychological block exists in the minds of many people as to the use of fuel around electrical machinery. Considering all of this, it was decided to develop this machine to use engine oil (military specification *MIL-L-7808* B) as a coolant.[16]

The decision to use oil as a coolant emphasizes the necessity of building a brushless machine. Several years ago experiments were run to determine the compatibility of military specification *JP-4* fuel with brushes. It was found, at that time, that brushes operating on slip rings would work satisfactorily submerged in fuel if the proper grade of brush

were used and the brush pressures properly adjusted. Similar tests on oils were far from encouraging. Neither *JP-4* fuel nor oil in droplet quantities was conducive to good brush operation.

When it was decided to use oil as a coolant, it was necessary to determine how it should be used. Should the machine be treated merely as an enlargement in a tube with all components in the machine submerged? Should the oil be sprayed on the hot components, then collected in a sump and pumped off? Or should the oil be put through separate ducts and the heat conducted to these ducts? Calculation based on viscosity and shearing in the oil indicated that a submerged system would absorb too much power. Tests confirmed this. Spray cooling would cause too much aeration of the oil, deterioration of the oil with overheating, and erosion of the insulation. It appeared most desirable to keep the oil pretty well contained in its own passageways. Tests on materials described later indicated that it was not desirable to have copious supplies of military specification *MIL-L-7808* oil around copper.

A machine cooled by oil but not having its internal parts submerged will be filled by various vapors. Since a machine of this type would of necessity be designed to operate at fairly high temperatures, some of the silicone compounds would probably be used. The vapors of these compounds are usually destructive to
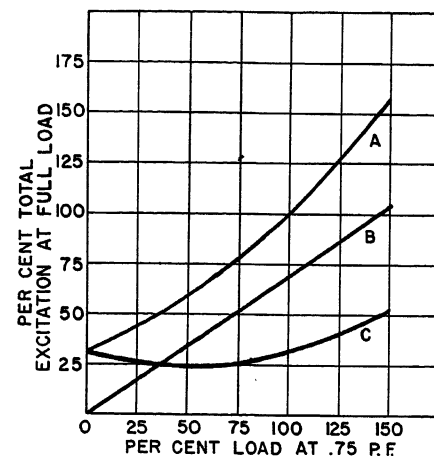
Fig. 7. Brushless a-c generator excitation requirements with constant generator output volts

Curve A—Total exciter requirements for rated voltage
Curve B—Excitation furnished by current transformers
Curve C—Excitation required from regulator

brushes.[6] Use of brushes in such a machine would rule out many otherwise usable materials.

Oil flow rate was established by a process of reasoning. Too large a flow rate would result in heavy, large ducts and too small a one would result in a large temperature rise in the oil. Small flow rates would make it difficult to set up turbulent flow in the ducts. Since oil has poor thermal conductivity, turbulence is a necessity for good heat transfer. Considering these things, a compromise of 5 gallons per minute was arrived at. This results in an oil rise through the a-c generator and a hydraulic drive of about 25 degrees centigrade (C).

To make the heat exchangers as light in weight and as small as possible, oil temperatures should be high. Above 180 C *MIL-L-7808 B* oil deteriorates rapidly. Considering this and the 25 C rise of the oil in passing through the generator and the drive, and allowing a slight margin, it was decided that design of the generator drive package should be based on an oil-in temperature of 150 C and the generator should use the oil before it is used by the drive.

In the authors' experience, static seals have been reliable if properly installed. Rotating seals have not been so reliable. Sometime during the life of a moving seal, it leaks, particularly, if it has to withstand pressure. At the temperatures just mentioned added trouble was expected.

A number of tests on rotating carbon type seals were run. Performance was good, but leakage of small amounts was
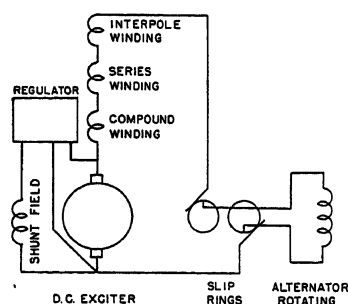
Fig. 5 (left). Schematic diagram of a conventional aircraft a-c generator
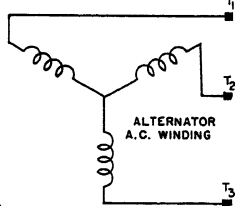
Fig. 6 (below). Schematic diagram of a brushless a-c generator, showing current transformer feedback and permanent magnet generator
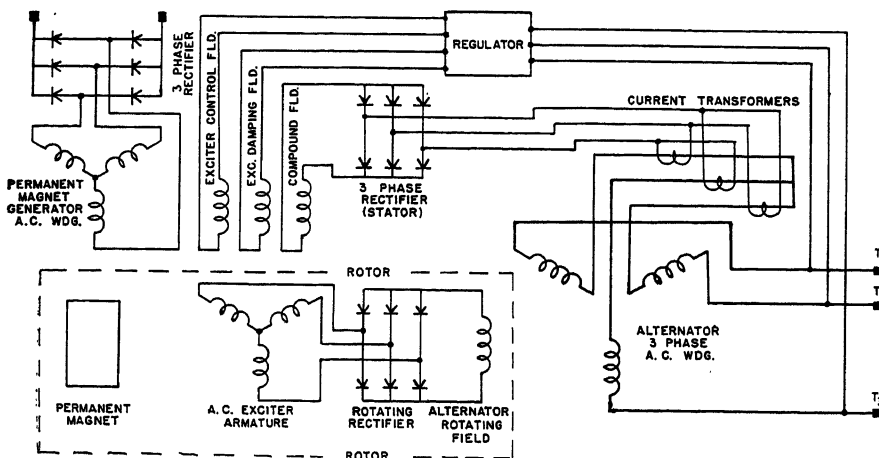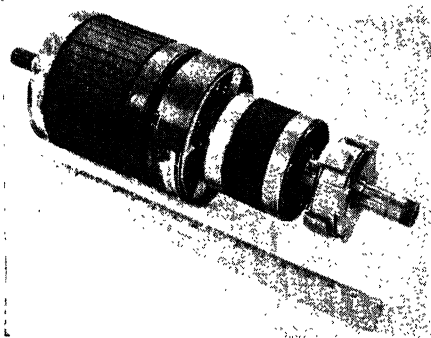
Fig. 8. Brushless a-c generator rotor

Fig. 9. 40-kva oil-cooled brushless a-c generator

encountered often enough to justify the following additional design criteria:

1. A sump line should be included and provision made to sump the generator and return to the oil system any oil that may leak by the seal during the life of the generator.

2. All materials used in the generator should be relatively impervious to *MIL-L-7808B* liquid or vapor.

Rumors received from many sources said that *MIL-L-7808 B* engine oil was an excellent solvent, paint, varnish and plating remover, and metal dissolver. To verify or disprove the action of the oil and to find satisfactory materials to build a machine over 100 different materials normally used in generator construction were submerged in the oil and heated, and held at temperature for 1 week. At the end of that time, many of the materials had dissolved in the oil. Others had become brittle. Some had changed their characteristics completely. Cadmium plating was removed from steel. Coppers and brasses, in general, grew thick crusts. Many silicone rubbers ceased to have rubber characteristics. Adhesives ceased to adhere. In general, glasses, teflon, asbestos, and mica were unaffected. Binders were frequently unsatisfactory.

Similar tests were run on the same materials except that the material was suspended in the vapor atmosphere above the oil. Results were different. Many materials which were unaffected by the liquid became useless in the vapor. The silicone rubbers showed the greatest variation in this respect.

In general, the basic materials, glass, asbestos, and mica, were not affected. Aluminum and magnesium appeared to be unaffected by either liquid or vapor. Copper with an insulation which would stand the liquid or vapor appeared to be not seriously affected. Fig. 4 shows a rack of tests containing a number of materials in oil. The vapor tests were run in the same way except that the mate-

rials were suspended above the oil.

Brushless machines may be made in many ways. Some of the more well known are inductor generators, cascade generators, permanent magnet generators, asynchronous generators, and generators with a-c exciters and rotating rectifiers in place of commutators and slip rings. Inductor generators have high leakage reactance and are inefficient in their use of iron in that the magnetic flux changes from some minimum plus value to some maximum plus value; hence, they are heavy for the rating. Cascade generators are generally slow in response and inefficient. Permanent-magnet generators are large and heavy for their ratings and their output voltage is not easily regulated. Asynchronous generators with shunt-capacitor banks have no short-circuit capacity. Series-capacitor banks overcome this objection but in either case, no adequate means for regulating asynchronous machine output voltage are available. This leaves the rotating-rectifier type machine as the most promising. It should have recovery and short-circuit characteristics similar to a-c generators used in aircraft power systems in the past, and a very similar type of regulating means should be applicable.

The design of an a-c generator must be co-ordinated with the excitation system to be used. Past a-c generators have consisted of a main machine with a static armature and a rotating d-c field fed by slip rings from a d-c exciter which has rotating armature and a static d-c field. Excitation for the exciter was obtained either from the output of the exciter or from the output of the main a-c generator. Stabilization feedback was obtained from the output of the exciter. Fig. 5 shows the schematic diagram of such a generator. Series excitation and armature reaction compensation in the exciter were obtained from the output of the exciter. In the rotating rectifier, many of these things must be obtained in different ways. Fig. 6 shows the schematic of a rotating-rectifier machine. In this unit, the main machine still has a static a-c armature and a rotating d-c field. Instead of rectifying the output of the exciter armature with a set of brushes and a commutator, the output is rectified by a rectifier which rotates with the shaft. The output of the rotating rectifier is fed directly into main machine rotating field. Series excitation and armature reaction compensation for the exciter can no longer be obtained from the output of the exciter. Instead, they are obtained by use of a current feedback system from the output
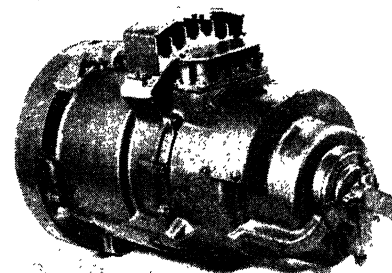
of the main a-c generator. Voltage control excitation is obtained from the voltage output of the main a-c generator. Fig. 7 shows test curves of the distribution of this excitation. Negative feedback for stabilization is obtained from a special damping winding on the exciter stator; Fig. 6. To insure stability of the system, all constants were calculated and put on an analogue computer where they were varied to give the time constants required for stable operation. The constants indicated by the analogue computer were incorporated into the machine and regulator design.

Rectifiers to be satisfactory for a rotating-rectifier machine cooled by a 150 C oil must fulfill certain obvious requirements. They must be capable of operating at some temperature in excess of 150 C. They must be capable of withstanding centrifugal force in the order of 15,000 g. They must be light, must occupy small volume, and must be sturdy, reliable, and preferably efficient. There are some less obvious requirements made necessary by the fact that transients reflect through a machine. To determine
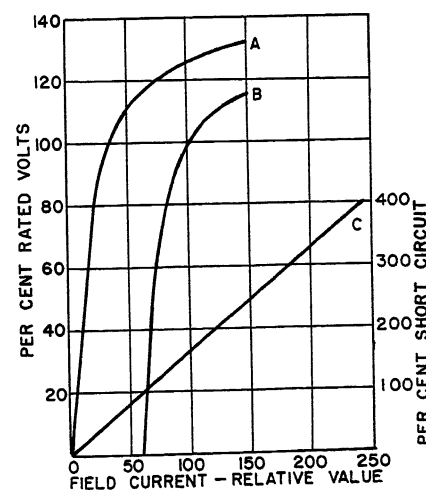
Fig. 10. Saturation curves

Curve A—No load saturation
Curve B—Full load saturation at 0.75 power factor
Curve C—Three-phase short-circuit saturation

by test how severe these requirements might be, a 40-kva air-cooled machine was set up with selenium rectifiers feeding its slip rings. Load-on, load-off and other transients were checked. It was found that the maximum back voltage reflected across the rectifier was approximately 50% higher than the maximum voltage out of the rectifier under any load condition. It was found that the maximum current in the forward direction was about twice the maximum current out of the rectifier for any load condition. These are somewhat generalized conclusions and would vary with the constants of the generator.

None of the known types of rectifiers, selenium, copper oxide, copper sulphide or germanium, would satisfy these requirements. However, early in 1954, the authors' company produced on a laboratory basis a rectifier cell based on the element silicon. This rectifier cell was capable of operating at 20 amperes in a single-phase bridge at 100 volts back voltage. Physicists forecast that this rectifier would be able to operate to temperatures as high as 200 C. Tests upon the cell indicated that when built into a rectifier, the efficiency would be as high as 95%. All of these things were necessary in order to build an adequate rotating-rectifier machine. To find out if these cells would operate under high centrifugal stress, would take short time voltage and current transients, and would operate at the temperatures required, test fixtures were built and each point was proved.

Conclusion

With the answers procured to all these questions, it was now possible to start the design of a machine based upon the answers. An oil ductwork system capable of taking 400 pounds per square inch was designed. This system first passes the oil down the center of the shaft in a stationary tube, then back along the inside diameter of the shaft, where it removes the rotor heat and the heat from the rectifiers. The oil is then passed through cast-in stainless steel tubing in the support castings where it picks up heat from the stator portions of the machine. One rotating seal is used. All other seals are static. All cooling calculations were made on the assumption that the heat must be conducted to the oil by the metallic or solid portions of the machine. On this basis, the main machine d-c rotor was designed to be of the round or distributed winding type to get better heat conduction paths without complex ducting.

The electrical design is straightforward, limited only by the maximum temperatures that can be tolerated and the maximum amount of heat that can be conducted to the oil with those temperatures. Of course, the compounding circuit must be balanced to give adequate short-circuit capacity and still leave regulation range. Fig. 8 shows the rotor of the machine. At the left is the d-c field of the main alternator. Next to it is the rotating-rectifier structure. The second rotor from the right is the exciter armature. At the far right is the permanent magnet field structure. Fig. 9 shows the assembled machine with a test bracket.

A considerable number of tests have been run on the machine. Fig. 10 shows saturation data for the main a-c generator obtained with special test fixtures. Fig. 7, shows tests data for the over-all machine excitation requirements. Harmonic analysis of the output indicates less than 2% for any one harmonic. Tests on unbalance indicate that the two-thirds requirement of *MIL-G-6099* has been met. Temperature runs and life runs, so far made, indicate that the original objectives, solution of the bearing, brush, and temperature problems, are attainable.

References

1. LUBRICATION LIFE FORECAST AND TEST PROCEDURE. *Engineering Manual*, General Motors Company, Bristol, Conn., May 1950.

2. DEVELOPMENT AND EVALUATION OF HIGH TEMPERATURE GREASES, E. A. Swakon, A. C. Borg. *Quarterly Report No. 13*, Wright Air Development Center, Dayton, Ohio, July 1954.

3. HIGH-ALTITUDE BRUSH PROBLEM, D. Ramadanoff, S. W. Glass. *AIEE Transactions (Electrical Engineering)*, vol. 63, Nov. 1944, pp. 825–30.

4. TREATMENT OF HIGH ALTITUDE BRUSHES BY APPLICATION OF METALLIC HALIDES. Howard M. Elsey. *AIEE Transactions (Electrical Engineering)*, vol. 64, Aug. 1945, pp. 576–9.

5. AIR HUMIDITY AND BRUSH CONTACT DROP: THE EFFECT OF WATER VAPOR, SULFUR DIOXIDE, AND HYDROGEN SULFIDE, Howard M. Elsey, Laurence E. Noberly, John L. Johnson. *AIEE Transactions*, vol. 73, pt. III-B, Dec. 1954, pp. 1383–89.

6. EFFECTS OF SILICONE VAPOR ON BRUSH WEAR, J. Marsden, Robert H. Savage. *AIEE Transactions*, vol. 67, pt. II, 1948, pp. 1186–90.

7. NEW TEST CHAMBERS FOR AIRCRAFT ELECTRIC APPARATUS WITH PARTICULAR REFERENCE TO CARBON BRUSHES, E. R. Summers, J. F. Settle. *AIEE Transactions*, vol. 63, 1944, pp. 1205–12.

8. THE TESTING OF BRUSHES FOR LIFE AND PERFORMANCE UNDER VARIOUS ALTITUDE CONDITIONS, C. J. Herman. *Ibid. (Electrical Engineering)*, Dec., pp. 929–33.

9. COOLING PROBLEMS OF ELECTRICAL EQUIPMENT IS HIGH-SPEED, HIGH ALTITUDE AIRCRAFT, Dr. E. Nauman. *Technical Data Digest*, Armed Services Technical Information Agency, Dayton, Ohio, vol. 17, no. 6, June 1952, pp. 28–36.

10. HIGH-TEMPERATURE PROBLEMS OF ROTATING ELECTRIC EQUIPMENT FOR AIRCRAFT, O. E. Buxton, Jr. *Electrical Engineering*, Jan. 1956, pp. 41–46.

11. PROCEEDINGS OF THE FIRST CONFERENCE ON COOLING OF AIRBORNE ELECTRONIC EQUIPMENT, *Bulletin No. 148*, Ohio State University, Columbus, Ohio.

12. MOTOR TESTS EVALUATE THERMAL ENDURANCE OF CLASS-H INSULATION AND SILICONE VARNISH, George Grant III, Graham Lee Moses, T. A. Kauppi, G. P. Gibson. *AIEE Transactions*, vol. 68, pt. II, 1949, pp. 1133–38.

13. GENERATORS AND REGULATORS; AIRCRAFT ALTERNATING CURRENT, GENERAL SPECIFICATION FOR. Military Specification *MIL-G-6099*, U. S. Bureau of Aeronautics, Washington, D. C., April 19, 1950.

14. LIQUID COOLING OF A-C TURBINE GENERATORS, Carl J. Fechheimer. *AIEE Transactions*, vol. 66, 1947, pp. 553–64.

15. LIQUID COOLING OF A-C TURBINE GENERATORS—PART II, C. J. Fechheimer. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 165–73.

16. LUBRICATING OIL. Military Specification *MIL-L-7808B*, U. S. Air Force, Washington, D. C.

◆

Discussion

D. W. Exner (Boeing Airplane Company, Seattle, Wash.): The significance of this major advance in the design of aircraft generators cannot be overstressed. It is an important step toward the development of truly environment-free equipment. As flight speeds increase, the cost in horsepower of blast air cooling becomes excessive. As speeds increase further, the stagnation temperature rise begins to affect the machine rating and then, finally, exceeds the allowable operating temperature of conventional insulating materials and lubricants. This is the cause of the trend toward use of fuel as a heat sink in advanced aircraft designs.

Direct use of fuel as a coolant, without interposing oil as an intermediate heat transfer fluid, will become attractive for certain applications. To gain maximum advantage from the use of fuel as a heat sink, further development of engine fuel systems is needed to permit us to use the heat of vaporization of the fuel. At present this presents problems in the metering and injection of fuel to the engine and in formation of coke deposits in the fuel system. It is believed that these problems can be solved by further research, thus making available a tremendously increased heat sink capacity.

H. J. Braun: The discussion of the direct use of fuel as a coolant is interesting to many electrical designers in that it indicates the trend of thinking in the aircraft industry. Fuel can be used as a liquid coolant or as an evaporant. Much of the fuels liquid cooling capacity is today wasted during loiter or let down conditions by the engine fuel-handling system. A considerable amount of development work needs to be done on variable displacement fuel pumps to alleviate this condition.

Mr. Exner discussed some of the problems encountered in vaporizing fuel. At the present time some engine manufacturers are working with means to handle fuel in the vapor form.

Air Turbine Drives: Performance and Limitations

L. ROYCE
NONMEMBER AIEE

THE practice in modern jet aircraft of using axial-flow jet engine compressors has decreased the number of engine pads available for driving accessories. This, coupled with the presence of a high-pressure air supply as a source of power, has brought about the introduction of the air turbine as a prime mover. Since the energy level of the jet engine bleed air is dependent upon the engine setting, which is in turn dependent upon the airplane flight condition, the design of the turbine drive will be dictated by, first, the flight condition which results in a minimum energy supply to the prime mover and, second, the power that must be developed by the prime mover. These factors will influence the performance of the unit at the airplane cruise condition where air consumption is most critical. The first section of this discussion is concerned with the imposed requirements at the design point, and the penalty involved from the standpoint of performance, at the cruise condition. Subsequent sections will deal with the turbine configuration and the turbine control system response rate.

Thermodynamic Performance

The horsepower (HP) developed by a turbine is proportional to the product of the temperature drop of the air passing through the turbine and the turbine air flow rate. Turbine temperature drop is used in place of enthalpy drop since the specific heat of air is assumed constant. The turbine air temperature drop depends upon the pressure and temperature of the air supplied to the unit at a given altitude and the turbine efficiency. The weight flow through the unit is dependent upon the pressure and temperature of the air supplied to the unit and the turbine nozzle area.

Figs. 1 and 2 represent a typical set of

bleed pressures and temperatures available from the jet engine compressor for the idle let-down and cruise flight conditions. With these values and with the assumption that turbine efficiency is 100 per cent (%), the maximum possible temperature drop through the turbine may be found; see Fig. 3, which shows that the turbine air temperature drop increases with altitude and the cruise temperature drop is the larger at any given altitude. Referring to Figs. 1 and 2, and using a nozzle area of 1 square inch, will show that the weight flow per square inch of nozzle area for each flight condition may be calculated. The results are plotted in Fig. 4, which shows that weight flow decreases with increasing altitude and at a given altitude is smaller for the idle let-down condition. Combining the results of Figs. 3 and 4 results in the HP that the unit can develop when the turbine efficiency is 100% and the nozzle area 1 square inch. Fig. 5 shows that the capacity of the turbine is always greater at cruise than at idle let-down and that the 10,000-foot idle let-down condition is the most critical operating point of the turbine. The HP developed at any operating point may be changed by selecting a new nozzle area; the new HP value being equal to the HP from Fig. 5 multiplied by the new nozzle area. Thus it can be seen that for any given nozzle area, operation at 10,000-foot idle let-down will always result in the smallest HP. Therefore, for the assumed bleed conditions, 10,000-foot idle let-down is the design point for the turbine.

Fig. 5 shows that if the turbine is sized to produce a given HP at the design point, then at all other operating conditions this same turbine will be capable of developing a greater HP. If it is desired to run the unit at constant load at various operating conditions, some method must be employed to limit the turbine power. Since HP is dependent upon turbine air temperature drop and weight flow, a control of either of these values will result in constant turbine power. Since turbine air temperature drop depends upon turbine supply pressure the turbine air temperature drop at any operating point may be decreased by

reducing the bleed air pressure at the inlet to the unit. This type of control is a throttling control using a constant or fixed turbine nozzle area. Control of turbine HP may also be accomplished by regulating the weight flow through the turbine. Since the weight flow is directly proportional to the turbine nozzle area, the nozzle area may be varied to vary the turbine HP. This configuration is called a variable area turbine.

The performance of a fixed- and a variable-area turbine is almost identical at the design point. At this operating point the supply conditions are the same for both units and so, for a given turbine HP, both units will have approximately the same nozzle area and thus the same air consumption. If operation of the two units at an off design point is now considered, it can be seen that the variable area unit will require a nozzle area decrease to maintain the same HP while the fixed area unit will require throttling. Throttling reduces the available energy supplied to the unit per pound of air and thus the fixed area unit will require a larger number of pounds of air to produce the required HP at the off design point than the variable area unit which is working with the higher energy air supply.

Assume that the turbine drives an alternator rated at 12 kva. If the power factor is 0.9, the alternator efficiency 0.8, and the turbine gear box efficiency 0.95, then rated turbine load is 19.1 HP. If the unit is designed so that the maximum power output at 10,000-foot idle let-down is 19.1 HP, then with 100% turbine efficiency the turbine required nozzle area would be 0.667 square inch. For a design point maximum load of 125% rated (23.9 turbine HP) the required nozzle area would be 0.835 square inch. Table I lists the nozzle area requirements for the 10,000-foot idle let-down design point for various load requirements.

The off design point of major importance is the point at which the most extended period of operation occurs. Fix this point at 40,000-foot cruise with the turbine operating at 100% rated load (19.1 HP). If it is assumed that the turbine efficiency is 100% at 40,000-foot cruise, the required air flow rate may be calculated to carry 100% load for a fixed nozzle area design with the use of each of the five design nozzle areas listed in Table I. The results of this calculation are shown by the higher weight flow curve of Fig. 6.

Inspection of Fig. 6 shows that as the load at the design point is increased, the air consumption required to enable the turbine to develop the same HP (100%
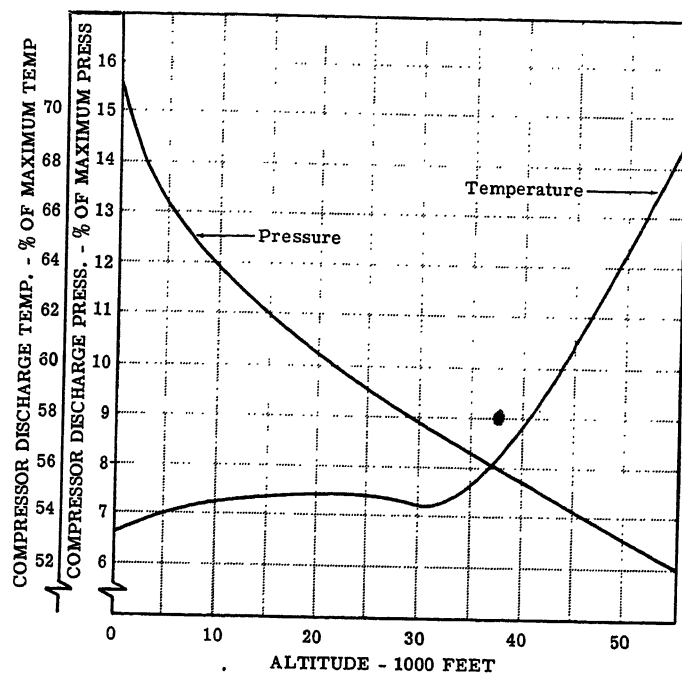
Fig. 1. Bleed pressure and temperature versus altitude-idle let-down
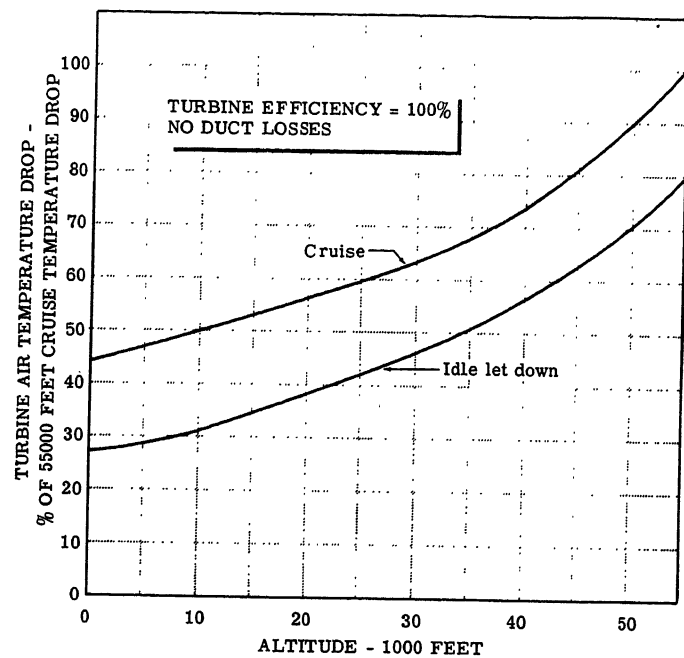
Fig. 3. Turbine air temperature drop versus altitude

rated) at 40,000-foot cruise increases. Since increasing the load at the design point increases the power capacity at the cruise point, Fig. 5, it is seen that each additional increase in load at the design point results in an additional increase in the amount of throttling required at the 40,000-foot cruise point to keep constant

load (100% rated). Hence the turbine operates with air at a lower energy level for each increase in throttling, and therefore the turbine requires a greater weight flow to develop the same HP.

For the five design points listed in Table I a variable area turbine operating at 40,000-foot cruise, 100% rated load, and

100% turbine efficiency would require the air consumption shown by the lower weight flow curve of Fig. 6, which shows that the air consumption for 100% load and 100% turbine efficiency at 40,000-foot cruise is independent of the load at the design point This is true because no throttling is present in a variable area design.

Fig. 6 shows the variation with load at the 10,000-foot idle let-down design point of the ratio of air consumption for a fixed nozzle area turbine operating at 40,000-
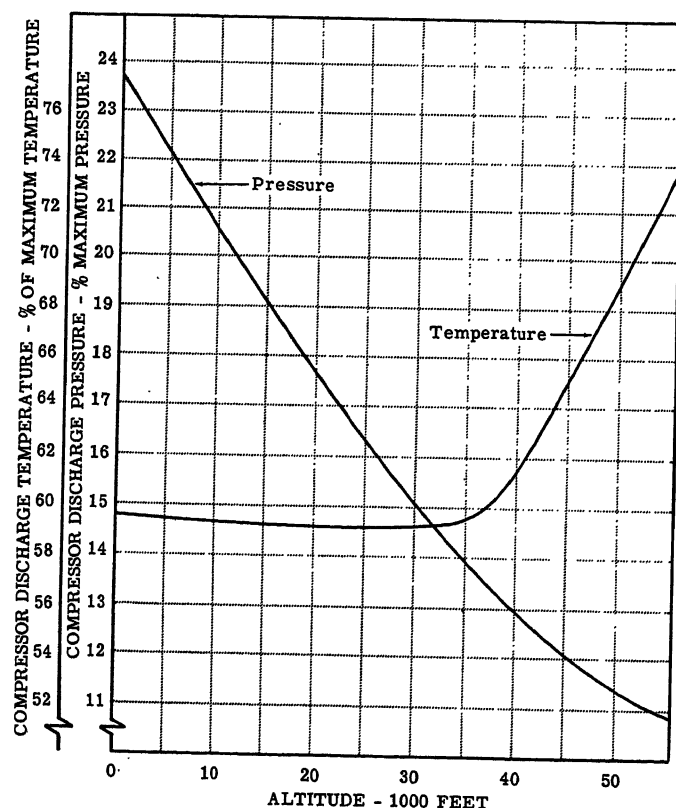
Fig. 2. Bleed pressure and temperature versus altitude cruise.
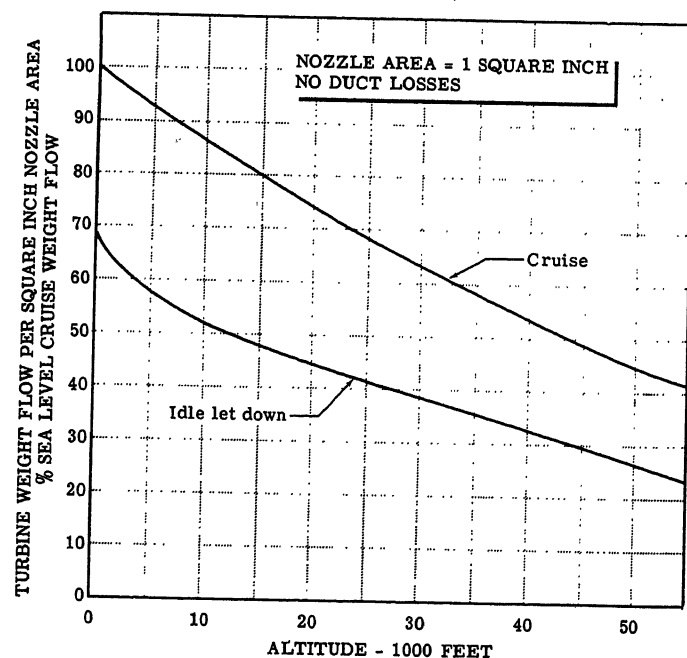
Fig. 4. Turbine weight flow versus altitude

Royce—Air Turbine Drives: Performance and Limitations
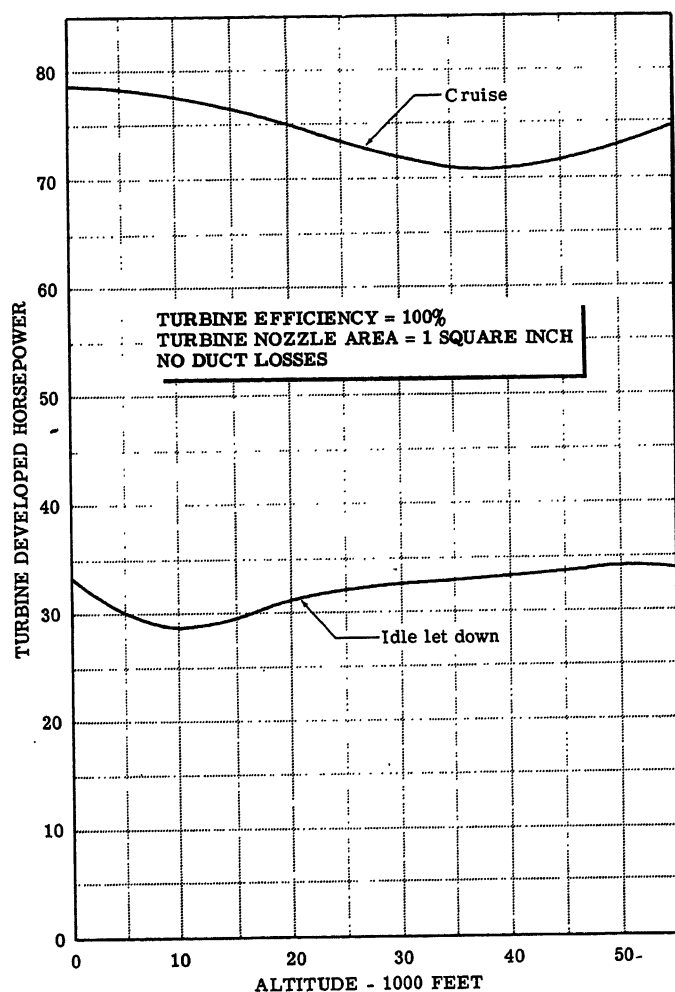
Fig. 5. Turbine developed HP versus altitude

Fig. 6. Weight flow at 40,000 feet, cruise 100% load versus per cent rated design load; ratio of fixed-area to variable-area weight flow versus per cent rated design load

foot cruise, 100% turbine efficiency, and 100% rated load to the air consumption for a variable area turbine operating at the same condition and same efficiency. These comparisons have been made on the basis of 100% turbine efficiency and assuming no pressure drop due to ducting between the jet compressor and the turbine.

With the duct characteristic curve shown in Fig. 7 and actual turbine efficiencies for fixed and variable area turbines the data shown in Fig. 6 may be recalculated.

Fig. 8 is a plot of the actual weight flow

Table I. Nozzle Area Requirements 10,000-Foot Idle Let-Down, 100% Turbine Efficiency

Load Requirement, Per Cent of Rated	Turbine, Required HP	Nozzle Area Required, Square Inches
100	19.1	0.667
125	23.9	0.835
150	28.6	1.0
175	33.4	1.17
200	38.2	1.335

required at 40,000-foot cruise, 100% rated load for both the fixed and variable area designs versus load at 10,000-foot idle let-down, and the ratio of the actual weight flows for the two types of turbines.

Three interesting observations can be made from Figs. 6 and 8. First, the air consumption for both types of turbines is greater when duct losses and actual turbine efficiencies are taken into account; second, the actual air consumption for the variable area design increases with increases in the load at the 10,000-foot idle let-down point; and, third, the per-cent increase in required weight flow for the fixed-area design compared to the variable area design is smaller in the actual case than in the ideal case of no duct losses and 100% turbine efficiency.

The first observation is almost self-evident. When actual losses are taken into account, the air consumption required increases. The fact that the actual air consumption of the variable area turbine increases at 40,000-foot cruise,
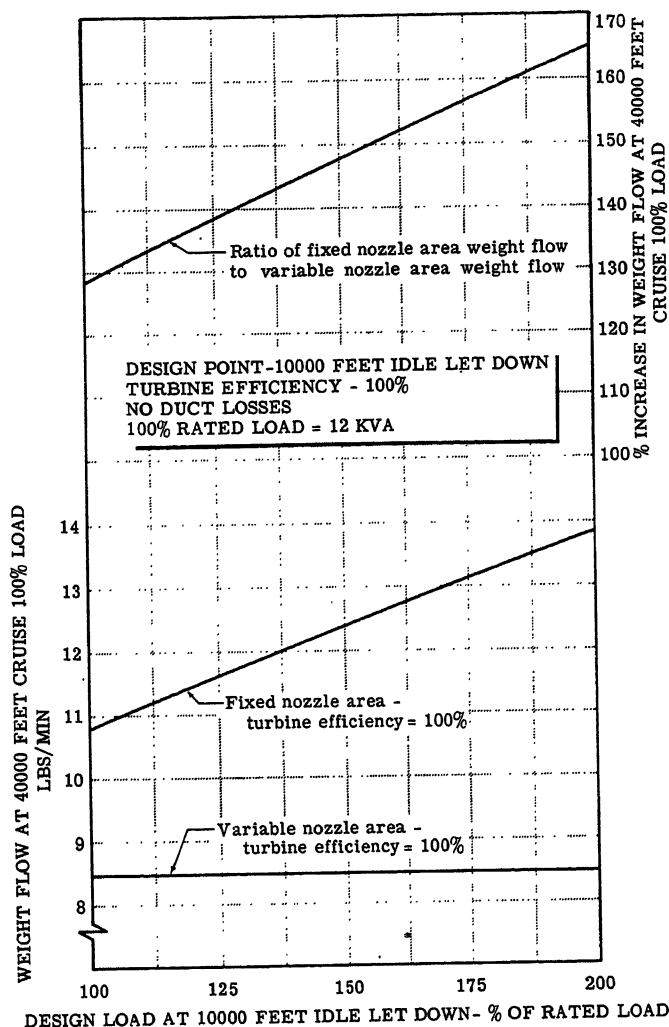
100% rated load with increasing load at the 10,000-foot idle let-down point (ideal air flow was shown to be independent of design point load) is due to the turbine efficiency variation with nozzle area. As the load at the design point is increased, the variable area maximum nozzle area increases, and so the nozzle area variation from the design point nozzle area to the 40,000-foot cruise, 100% load nozzle area increases and turbine efficiency falls.

The third observation is the most important. Here it is seen that the penalty of using a fixed area design rather than a variable area design does not result in as large a percentage increase in air consumption as would be indicated theoretically.

Fig. 8, based on typical bleed conditions and typical duct and turbine characteristics, shows the penalty which must be paid in air consumption at the cruise condition for increasing the load requirement at the design point. This figure shows that regardless of the type of tur-
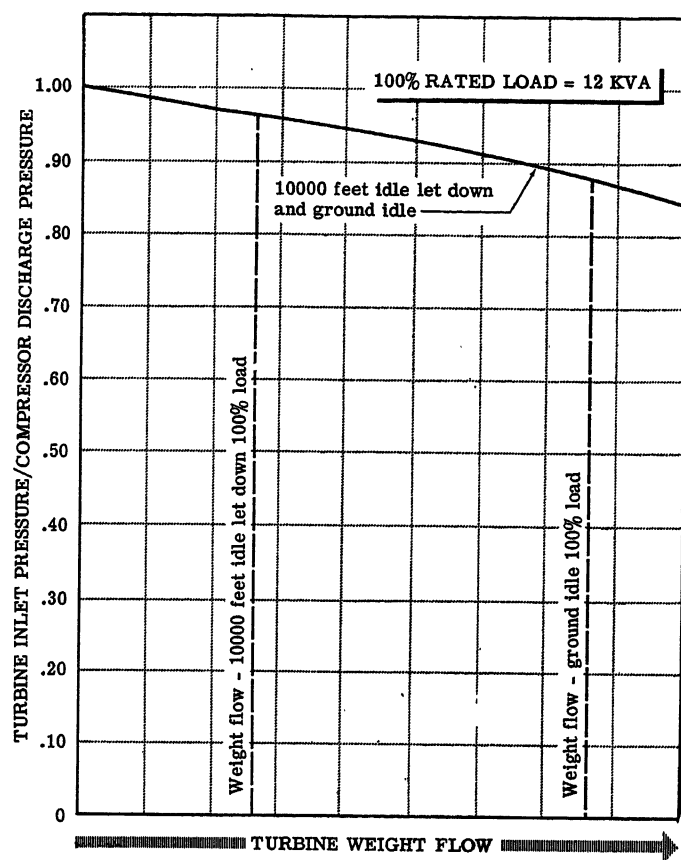
Fig. 7. Turbine inlet pressure/compressor discharge pressure versus turbine weight flow

Fig. 8. Actual weight flow at 40,000-foot cruise, 100% load versus per cent rated design load; ratio of fixed area to variable area, weight flow versus per cent rated design load

bine selected the weight flow will increase as turbine design load increases. By doubling the design load the fixed area unit operating at 100% load at 40,000-foot cruise will increase its air consumption by 36% while the variable area unit will increase its air consumption by 19%. It becomes imperative, then, to evaluate thoroughly the actual electrical loads at the design point if a realistic design load for the turbine is to be established. It should be remembered that the turbine will be capable of developing a greater HP than the design HP at all operating conditions other than the design point.

This analysis was carried out with the use of bleed data during flight. There are cases where the turbine is required to operate during a ground idle condition, in which case the design point for the turbine becomes ground idle. The ground idle design point is usually so much more critical to unit performance at the cruise condition that a short discussion is warranted to stress the importance of the load requirement at the design point.

Assume that the jet engine compressor ground idle pressure ratio is 1.6, making the bleed pressure 47.8-inch Hg. Further assume the same duct characteristic as was used for the 10,000-foot idle let-down design point; see Fig. 7. The HP

output of the turbine for the ground idle condition may now be plotted as a function of the turbine nozzle area. Fig. 9 shows that for the bleed pressure of 47.8-inch Hg and the assumed duct losses, the actual turbine HP and the turbine HP based on 100%, turbine efficiency both increase as larger values of nozzle area are selected up to a certain value of nozzle area. Selecting still larger nozzle areas results in a decrease in turbine power rather than an increase. This is a very important characteristic because it shows that if the duct is too small, the duct actually limits the amount of power that the turbine can develop. Fig. 9 shows that 100% rated power (19.1 HP) can be obtained at ground idle but 125% is not possible with the assumed duct. For the 100% load at ground idle as design point, the actual air consumption at 40,000-foot cruise, 100% load for a fixed area turbine would be 24.7 pounds per minute, representing an increase of 72% over the weight flow at 40,000-foot cruise, 100% load for a fixed area turbine based on a 100% load at the 10,000-foot idle
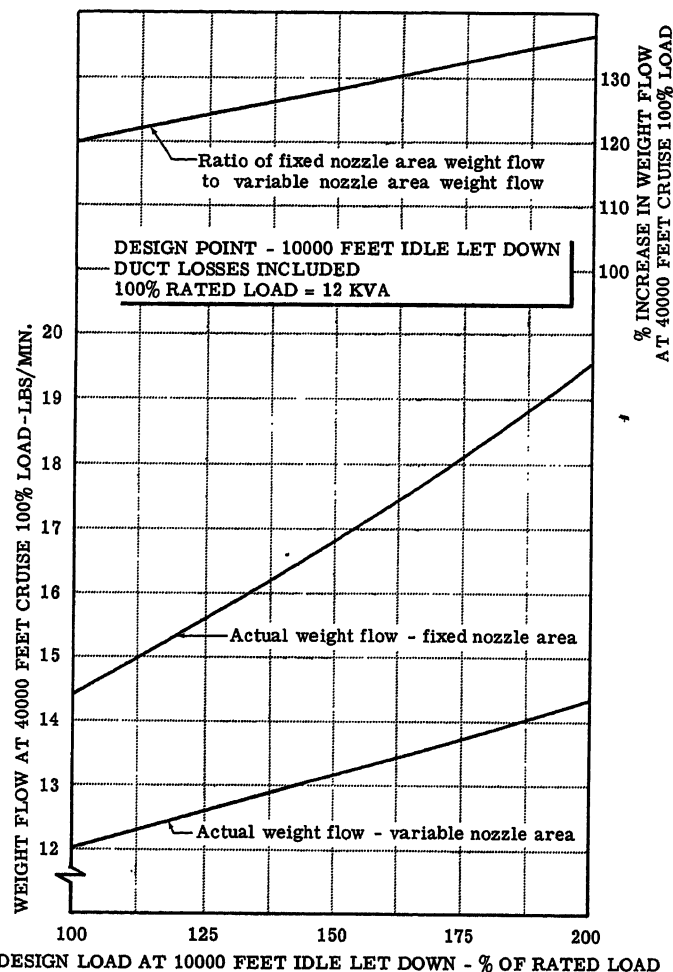
let-down design point. The 24.7 pounds-per-minute air consumption is plotted in Fig. 10.

The limitation in the turbine power at the ground idle design point is due only to the duct loss. Evidently the duct size may be increased, up to a point, and whatever power is desired obtained. However, the large increase in required weight flow at cruise is not due entirely to the duct loss. This can be seen by assuming no duct loss and using actual turbine efficiencies to calculate the 40,000-foot cruise, fixed-nozzle area, 100% load weight flows. These data are plotted in Fig. 10, and by comparison to Fig. 8 show increases in actual weight flow of approximately 39% over the weight flow at 40,000-foot cruise, fixed-nozzle area, 100% load, when the design point was 10,000-foot idle let-down and duct losses were included. The introduction of the ground idle point as the design point of the unit works a hardship on the variable-area as well as on the fixed-area design. Since the maximum nozzle area requirement increases in going from the 10,000-foot

Fig. 9 (left). Turbine-developed HP, versus nozzle area
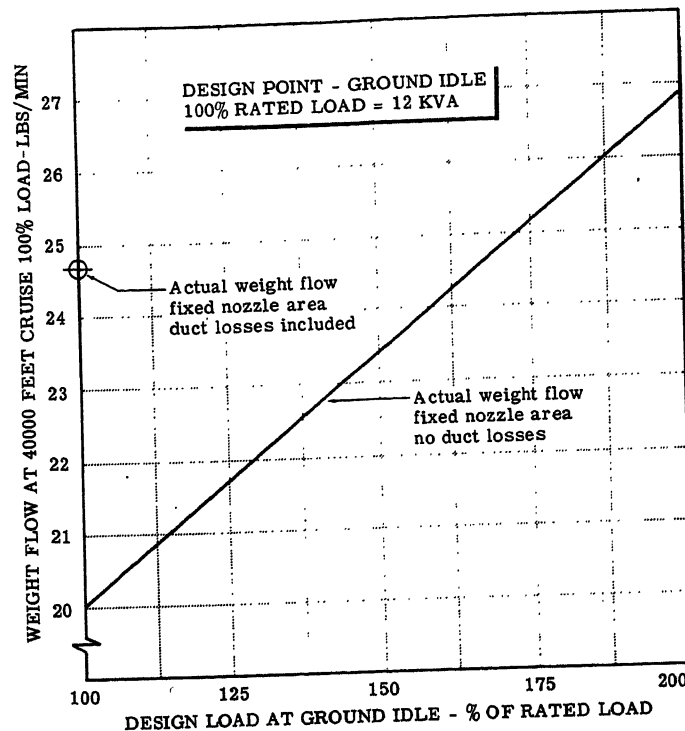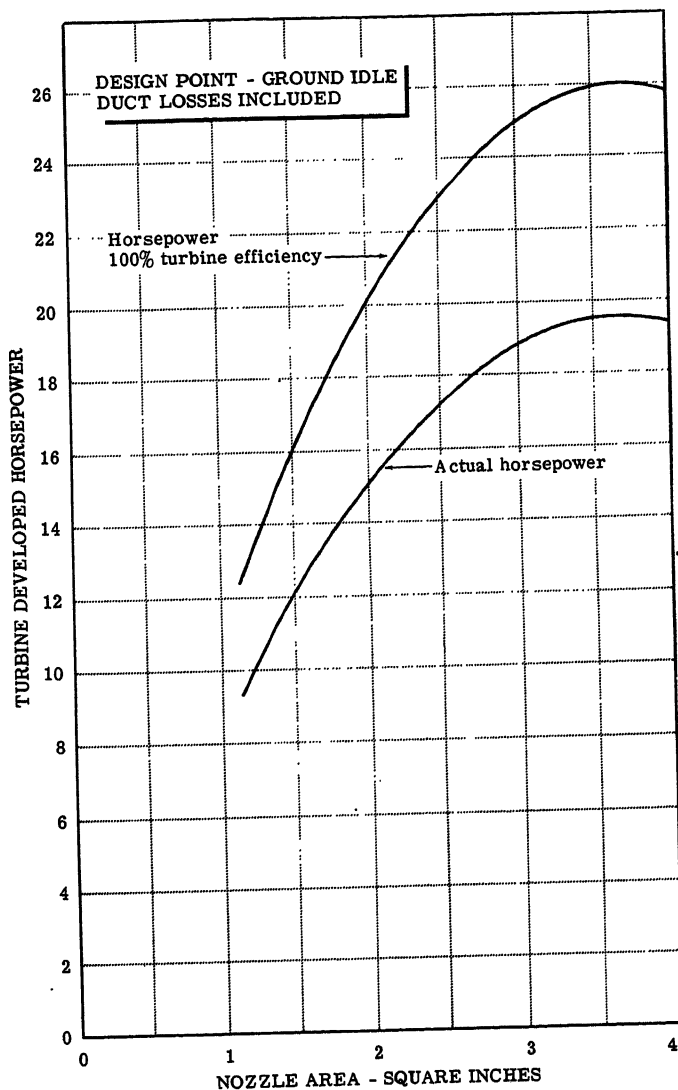
Fig. 10 (above). Actual weight flow at 40,000-foot cruise 100% load versus per cent rated design load

idle let-down to the ground idle design point the area ratio range has been increased through which the nozzles must be varied. There is a mechanical limitation to this variation and if the limit is exceeded, throttling must be employed. Therefore, without going into details of variable-area performance at cruise based on ground idle design, it is sufficient to say that the air consumption will be considerably in excess of the values shown in Fig. 8 for the 10,000-foot idle let-down design point.

The discussion has been confined to the performance of fixed- and variable-nozzle area turbines. In particular applications it may be advantageous to use designs involving partial admission, multiwheels, or variable turbine speed (in conjunction with constant output speed). The partial-admission arrangement is really a modified variable-area design where the nozzle area is increased in large steps rather than continuously. Multiwheel configurations are useful when the required operating range calls for two entirely different nozzle areas

and the losses involved in a partial-admission design are objectionable. Variable turbine speed units are helpful in enabling a better matching of the turbine characteristic to the bleed air pressure schedule. The thermodynamic performance of all of these turbines falls between the performance of the fixed- and variable-area designs and is adversely affected by increased load requirements at the design point.

It may be seen, then, that regardless of the type of unit used, the selection of the minimum acceptable load at the minimum jet bleed air energy point will afford considerable savings in unit air consumption at all operating points and particularly at cruise. In cases where duct sizes are limited by the unit location, it will allow a practical solution to a design of an alternator drive. If the turbine cannot develop the required power when load requirements are held to a minimum and duct size is a maximum, the last resort is to increase the idle speed of the engine to obtain the required turbine power.

Turbine Configuration

In general, there are any number of possible turbine configurations which may be used to develop power. The following discussion deals only with those turbines that experience has shown to be most adaptable to aircraft use.

The two basic types of turbines are the fixed-area and the variable-area. Each of these types can be broken down into the following classification based upon
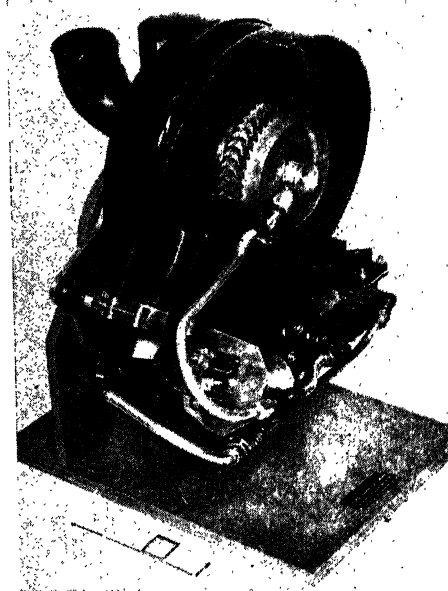
Fig. 11. Stratos model TP15-2, 9-kva power turbine

direction of flow of the air through the turbine. 1. Fixed area: axial flow; centripetal flow. 2. Variable area: centripetal flow.

The general physical dimensions of an axial turbine consist of small outside diameter and long length. The centripetal turbine is usually large in outside diameter and short in the length dimension. Comparing the weight and volume of a fixed- to a variable-area turbine shows that the fixed area unit is the smaller and lighter of the two. Also, the fixed-area turbine is much simpler mechanically.

The selection of the type of turbine to be used in a given application depends upon the load requirement, the allowable unit weight and volume, and the reliability required. In deciding whether fixed or variable area should be used the first consideration is the space and weight limitations. In many cases there is no choice since a variable area unit would not fit within the allowable envelope. However, if a choice exists, the next consideration is the magnitude of the load. If loads are small, although the fixed-area unit may require an appreciable percentage increase in air consumption over a variable-area design, this appreciable percentage may be a relatively small amount of air flow in terms of pounds per minute. In this case, it would not seem desirable to increase the weight and complexity of the turbine to save a small amount of air flow. (It should also be remembered that unit weight decreases airplane range just as does turbine air consumption, but to a smaller degree.) If, on the other hand, the load requirements are large, it may be desirable to use variable area with its associated complexity and larger weight in order to take advantage of large-weight flow savings.

In alternator drive applications, failure of the speed control system to hold speed as designed is as bad, from the standpoint of performance, as a complete loss of the unit. The variable-area unit, with each nozzle blade pivotable, is susceptible to much more mechanical difficulty than a fixed-nozzle design and thus susceptible to more speed control failures. This is an important characteristic of the variable-area unit and must be given due consideration in the selection of the type of turbine to be used in a given application.

Fig. 11 shows a cutaway of a power turbine. This unit uses an axial-flow design and is geared to drive a 4,000-rpm 9-kva alternator.

The turbine drive configuration that is used in any particular installation is usually a unit compromised with respect to the many controlling requirements.

It must be remembered that in the overall picture the airplane performance rather than the unit performance is the governing criterion.

Turbine Speed Control Response Rate

The nature of the electrical loads found in today's airplanes requires that the speed of the alternator be held within relatively close limits during steady-state and transient operation. The means by which the speed control requirements may be met are numerous. Electric, hydraulic, and pneumatic systems or any combination of these are possible. The response rate of a given system will depend upon the design of the system. As an example of the state of the art at this time, the response of the Stratos pneumatic speed control system is presented. This control system is isochronous and is used in conjunction with a fixed-nozzle area turbine.

STEADY STATE RESPONSE

The speed control functions as an isochronous controller within ±0.50% of steady-state speed. Setting of the speed control is held to within ±1.5% of rated speed.

TRANSIENT RESPONSE

When step load increments of ±50% rated load are added or removed from the load, the speed recovers to 2.5% of the initial steady-state speed within 0.6 second and to 1.5% within 1 second. Under this condition the maximum deviation from the steady-state speed is not more than 4.5%.

The ability of the unit to return rapidly to a steady-state speed after the system is disturbed by load changes is an important consideration. It is customary to specify the speed error, after a given interval of time from the initiation of the disturbance, in the form of a percentage speed variation. However, it is important to refer this percentage to the proper speed. For the proportional type of governing systems the percentage speed variation after a given time interval is usually referred to the new steady-state speed. For an isochronous control system the steady-state speed is independent of the load. Therefore, to put the two types of control systems on a comparable basis, all percentages should be referred to the steady-state speed before the load change.

The response rate of a speed control system used with a variable-area turbine will be poorer than the response rate of

the system when used with a fixed-nozzle design. This is true because the inertia forces involved in positioning the variable nozzles are greater than those for positioning the throttle valve. The response of the variable area unit may be made faster by inserting a throttle valve upstream of the variable nozzles and positioning this valve as a function of unit speed. The variable nozzles could then be positioned as a function of flight condition and altitude. This would result in the same speed response as for the fixed-nozzle design but the throttle valve would introduce a slight throttling loss at all operating conditions.

Conclusions

The selection of the type of turbine to be used for a given application depends upon the requirements for the particular job. In general, a variable-area turbine will require less air flow, will weigh more, take more space, and be more complicated than a fixed-area unit. The performance of either type of unit at all operating points, and particularly at the cruise condition of the airplane, suffers as the load requirement at the unit design point is increased. Therefore every effort should be made to evaluate realistically the minimum acceptable load at the turbine design point, which usually occurs at the minimum energy level of the bleed air.

———————◆———————

Discussion

J. Bendersky (General Electric Company, Lynn, Mass.): Mr. Royce's paper is a great contribution to the new field of jet aircraft air turbine drives. This paper goes a long way in fulfilling the increasing need of shedding light upon the design problems of air turbine drives.

The author's point of the importance of not overestimating the load requirements at the low bleed air energy design point can not be overemphasized. The writer's company has spent considerable time and effort in studying the problem of minimizing the bleed airflow consumption under varying loads and varying bleed air energies.

We have found that the load requirements of our turbo alternator drives and turbo hydraulic pumps could vary by a factor of 100 to 1. This coupled with a bleed pressure variation of 250 pounds per square inch absolute (psia) at maximum energy to 8 psia at minimum energy and with a turbine exhaust pressure variation of 14.7 psia at sea level to 1.3 psia at altitude would lead to a desired turbine nozzle area variation of 3,000–5,000 to 1.

Obviously, a turbine nozzle area variation of 5,000 to 1 cannot be accomplished with-

out a considerable loss in turbine efficiency. Therefore, it is necessary to keep the required area variation to a minimum.

It can now easily be seen why there are mechanical difficulties in designing a variable area nozzle to cover such a large variation in area. For example consider a variable area nozzle which changes its area by rotation of the nozzle blades position from full closed through 20 degrees of arc to open. To reduce the full area by a factor of 5,000 the blades would have to be positioned 15 seconds of arc from the full closed position. This is impractical from a mechanical design standpoint because tolerances would allow some of the blades to be completely closed and some of the blades would be 0.5 degree open. In addition, the turbine efficiency would be very poor because of the small nozzle angle and because the blade tip leakage area would be many times as great as the passage area.

The throttling process which may simply be a butterfly valve in the bleed air supply line does not suffer from the mechanical problems of area variation. The butterfly valve is quite capable of reducing the turbine nozzle effective area from full to zero, an infinite variation, without any mechanical difficulty. The turbine efficiency at a high degree of throttling would not be lower than the variable area nozzle turbine efficiency at small nozzle areas and in fact might be higher.

D. W. Exner (Boeing Airplane Company, Seattle, Wash.): The author points up very well one of the problems of applying air turbine drives to electric generators, the problem of matching the design of the turbine to the design of the generator. This problem is common to all other prime-mover drives of limited capacity, such as auxiliary gas turbines and reciprocating engines.

The power system designer is faced with the problem of providing adequate electric power, not only under normal steady-state conditions, but also under transient overload conditions; also, he must be sure that sufficient power is available to clear circuit faults quickly. Failure to do the latter may result in complete loss of power, excessive damage, or fire, or at least a serious outage. Faults because of their serious potential effects, must be anticipated at any time and, because they may be caused by other airplane damage, they may occur coincident with conditions causing low air supply.

Electric a-c generator output is limited primarily by thermal and excitation capacities. Aircraft constant-frequency generators are normally designed to supply 150% load for 5 minutes and 200% for 5 seconds, and to deliver a minimum of 300% current to a solid 3-phase fault at their terminals. With certain types of faults and certain fault impedances the torque reflected into the prime mover by the generator will exceed 200% rated. The system designer must assume the worst fault condition because he has no control over the occurrence.

Because the small prime mover is inherently torque limited, the application problem becomes complicated. The inevitable result is pressure by the prime-mover designer on the system designer to compromise his normal application margins so as to lower the prime-mover design point. This pressure is good if the net result is a realistic check on system requirements. It is bad, on the other hand, if the over-all system reliability and safety is compromised.

As a matter of interest, this difficulty does not seem to arise in applying a hydromechanical transmission to a generator. This transmission does not have the same torque limitations on design as a small prime mover, and the extractable torque from the main engine is, for all practical purposes, unlimited.

The Effect of Machine Impedances on the Voltage Unbalance of 3-Phase Synchronous Generators

T. F. HARDMAN
ASSOCIATE MEMBER AIEE

IN COMMON use today are many 3-phase 4-wire electric systems that operate primarily with single-phase line-to-neutral loads. Quite often on these systems it is found impractical to distribute the single-phase loads so that the power supply operates into a balanced 3-phase system. A-c power systems on present-day aircraft typify this condition.

Servomechanism systems, radar apparatus, electric rate-gyros, and other diversified single-phase loads on aircraft

systems require close voltage regulation limits for satisfactory operation. It is imperative then that the aircraft power supply have characteristics which limit the deviation of the single-phase terminal voltages when unequal loading is applied.

Previous papers and texts have emphasized that the negative-sequence reactance reflects the ability of a 3-phase synchronous generator to prevent the unbalancing of the terminal voltages under conditions of unbalanced loading. This statement applies to systems on which the load unbalance is not extensive or where the other impedance quantities of the synchronous generator are small in comparison with the negative-sequence reactance. If the latter conditions are not true, further consideration must be given. Also lacking in previous work, without considerable calculations, is a working knowledge of the extent that the generator impedances, including the negative-sequence reactance, effect voltage unbalance.

Hence, the need arises for practical information that will enable the design engineer to evaluate the effect of the machine impedances on this aspect of performance in the consideration of the over-all performance of an optimum synchronous generator for a given application. Information of this nature would also permit the utilizer of the generator to make performance predictions from tested or calculated impedance values.

Background

The effect of applications of unbalanced loading on wye-wye neutral-connected systems has been treated in recent papers.[1,2] Prior to this time published work has been concerned mainly with the fault-type unbalance with little reference to the effect of the voltage regulator. To augment previous work, and because of its widespread application, the wye-connected regulated generator is chosen for this investigation.

The electric systems employed in aircraft are numerous, and the nature of the unbalanced loading which can occur on these systems is extensive and diversified. To present any analysis of performance it is necessary to select conditions which are defined and which impose the most severe requirements on the synchronous generator. The military specifications[3,4]
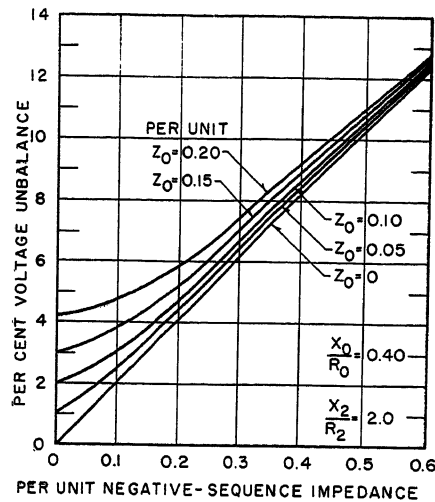
Fig. 1. The effect of the negative-sequence and zero-sequence impedances on the voltage unbalance of a synchronous generator carrying 2/3-rated line current on one phase only

Fig. 2. The effect of the negative-sequence and zero-sequence impedances on the voltage unbalance of a synchronous generator carrying 1/3-rated line current on one phase only
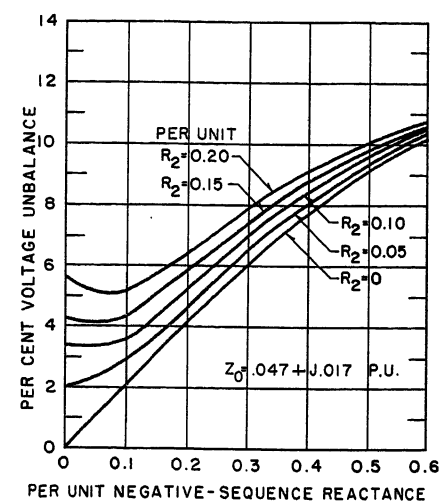
Fig. 3. The effect of the negative-sequence reactance and resistance on the voltage unbalance of a synchronous generator carrying 2/3-rated line current on one phase only

establish tests which have become commonplace as a measurement of the unbalanced voltage performance of this type of equipment. Since these test conditions do not represent probable application conditions and are primarily for the purpose of establishing generator characteristics, the most practical and best-defined case is when the generator is loaded on one phase only.

Nomenclature

a = phasor operator $\epsilon^{j(2\pi/3)}$
E = positive-sequence generated voltage
E_a, E_b, E_c = positive-sequence generated phase voltages
I = load current
I_a, I_b, I_c = phase currents
I_{a1}, I_{b1}, I_{c1} = positive-sequence phase currents
I_{a2}, I_{b2}, I_{c2} = negative-sequence phase currents
I_{a0}, I_{b0}, I_{c0} = zero-sequence phase currents
V_a, V_b, V_c = phase voltages
V_{a1}, V_{b1}, V_{c1} = positive-sequence phase voltages
V_{a2}, V_{b2}, V_{c2} = negative-sequence phase voltages
V_{a0}, V_{b0}, V_{c0} = zero-sequence phase voltages
Z_1, Z_2, Z_0, = positive-, negative-, and zero-sequence generator impedances
Z_{a1}, Z_{b1}, Z_{c1} = positive-sequence generator impedances
Z_{a2}, Z_{b2}, Z_{c2} = negative-sequence generator impedances
Z_{a0}, Z_{b0}, Z_{c0} = zero-sequence generator impedances
Z_L = load impedance

Method of Calculation of the Phase Voltages

The steady-state operating conditions of an unbalanced, regulated generator are a function of the generator excitation.

Knowledge of the effect that the regulator-sensing circuit has on the generator excitation permits a great simplification in the analytical solution for the generator terminal voltages. The solution as well as the understanding of unbalanced voltage conditions is made possible by the method of symmetrical components. The equations for the calculation of the terminal voltages of a 3-phase synchronous generator, with one loaded phase only, are developed in the Appendix and are

V_a = reference terminal voltage,

the loaded phase   (1)

$$V_b = a^2 V_a + \frac{I}{3}[(a^2-1)Z_0 + (a^2-a)Z_2] \quad (2)$$

$$V_c = a V_a + \frac{I}{3}[(a-1)Z_0 + (a-a^2)Z_2] \quad (3)$$

To calculate the terminal voltages from equations 1, 2, and 3, V_a is chosen as the reference voltage and is assumed to be some arbitrary value, say $120|\underline{0}$ volts. V_b and V_c are then calculated for a given magnitude and phase angle of the line current I. The magnitudes of the calculated values of V_a, V_b, and V_c are then re-evaluated by the ratio of the sum of the calculated values $|V_a|$, $|V_b|$, and $|V_c|$ to the adjusted regulated sum of $|V_a|$, $|V_b|$, and $|V_c|$.

Example 1

The problem is to find the value of the line-to-neutral voltages for a 15-kva 120/208-volt, a-c synchronous generator carrying rated 41.7-ampere line current at unity power factor on one phase only (3-phase average sensing), where

$Z_2 = 0.201 + j0.355 = 0.407|\underline{60.4}$ ohms

$Z_0 = 0.127 + j0.050 = 0.136|\underline{21.5}$ ohms

The solution is as follows: assume that

$V_a = 120|\underline{0}$

$V_b = a^2 V_a + \frac{I}{3}[(a^2-1)Z_0 + (a^2-a)Z_2]$

$V_b = 1|\underline{240} \times 120 |\underline{0} + \frac{41.7}{3}|\underline{0}[\sqrt{3}|\underline{-150} \times$

$0.136|\underline{21.5} + \sqrt{3}|\underline{-90} \times 0.407|\underline{60.4}]$

$V_b = 123.6|\underline{-115.7}$

$V_c = a V_a + \frac{I}{3}[(a-1)Z_0 + (a-a^2)Z_2]$

$V_c = 1|\underline{120} \times 120|\underline{0} + \frac{41.7}{3}|\underline{0}[\sqrt{3}|\underline{150} \times$

$0.136|\underline{21.5} + \sqrt{3}|\underline{90} \times 0.407|\underline{60.4}]$

$V_c = 130.5|\underline{123.4}$

Calculated

$|V_a| + |V_b| + |V_c| = 120 + 123.6 + 130.5$
$= 374.1$

Regulated

$|V_a| + |V_b| + |V_c| = 360$

Ratio

$\frac{360}{374.1} = 0.962$

Re-evaluated

$V_a = 120 \times 0.962 = 115.4|\underline{0}$

$V_b = 123.6 \times 0.962 = 118.9|\underline{-115.6}$

$V_c = 130.5 \times 0.962 = 125.5|\underline{123.3}$

This simplified method of solution is based on the assumption that the voltage regulator maintains the terminal voltage at a constant average value. The application of unbalanced loads on a regulated system creates a change in phase angle of the voltage sensed by the regulator and results in a departure from this average value. To determine the extent of this departure, the sums of the phase voltages were compiled for three different types of aircraft a-c generators operating under all the conditions of the military specifications.[3,4] The sums of the phase voltages varied from 358.4 to 362.0 with the regulator adjusted to hold the sum to 360. The largest variations occurred when the phase angles of the voltages had the most extreme variation from the balanced position. This compilation demonstrates the accuracy of calculation by this method. Under the most severe conditions of the military specifications,[3,4] the calculated values will be within 0.55 per cent of the terminal voltages that could be calculated by an exact method. The tested aircraft generators have a negative-sequence impedance of 0.50 per unit or less and a zero-sequence impedance of 0.05 or less. The error in calculation will increase with increasing sequence impedances and with more rigorous unbalanced load conditions.

The purposes for the development of equations in this form are as follows:

1. The equations are independent of the generator positive-sequence impedance Z_1 and the load impedance Z_L.

2. The magnitude and phase angle of the line current I can be positively and purposely defined.

3. A simplified solution of the regulated values of V_a, V_b, and V_c which gives reasonably accurate results can be obtained without the complication of the solution for these voltages by considering the circuit composed of the regulator, exciter, and alternator.

The absence of Z_1 in the equations simplifies the solution, emphasizes the sequence impedances which affect the unbalance on the unloaded phases, and permits an investigation of the variation of Z_2 and Z_0 without consideration of the change in Z_1.

The Effect of the Machine Impedances on Voltage Unbalance

The generator positive-sequence impedance has no effect on the unbalancing of the phase voltages of a 3-phase a-c generator. As shown by equations 29, 30, and 31 of the Appendix, the internal voltages caused by the positive-sequence impedance add in a like manner to each of the generated phase voltages, therefore maintaining balanced terminal conditions.

The effect of the variation of negative-sequence and zero-sequence impedance on the voltage unbalance performance of a regulated, 3-phase a-c generator carrying 2/3-rated live current on one phase is shown by the curves of Fig. 1, and for 1/3-rated line current on one phase by the curves of Fig. 2. These curves were constructed by computing the per-cent voltage unbalance, as defined by the military specifications,[3,4] from calculated values of phase voltages by the method of equations 1, 2, and 3. Various values of Z_2 and Z_0 were chosen; R_2 was allowed to vary proportionately with X_2 and R_0 to vary proportionately with X_0. The impedance values of the reference generator of example 1 were used as a base.

As shown by the curves of Figs. 1 and 2, the voltage unbalance of the synchronous generator is primarily determined by the negative-sequence impedance. However, as the value of the negative-sequence impedance is decreased to meet performance requirements, the zero-sequence impedance becomes increasingly more effective. Since a single-phase, line-to-neutral unity power-factor load is representative of the most severe specification test requirements imposed on the generator, these curves will assist the machine designer in the selection of generator windings and dimensions to obtain an optimum design for a given application.

The effect of the variation of negative-sequence reactance and resistance on the voltage unbalance performance of a regulated, 3-phase a-c generator carrying 2/3 rated line current on one phase is shown by the curves of Fig. 3. These curves were constructed by computing the per-cent voltage unbalance from calculated values of phase voltages by the method of equations 1, 2, and 3. Various values of Z_2 were chosen; R_2 was allowed to vary proportionately with X_2, with the use of the negative-sequence impedance value of the reference generator of example 1 as a base, and Z_0 was held constant at a value of $0.047+j0.17$ per unit.

It has been stated[5] that the negative-sequence reactance reflects the ability of the generator to prevent the unbalancing of the voltages, and the negative-sequence resistance reflects the ability to carry the negative-sequence current without undue rotor heating. As shown in Fig. 3, this statement is only partially true. The negative-sequence resistance also reflects the ability of the generator to prevent unbalancing of the voltages, especially as the value of negative-sequence reactance is reduced. These curves will assist the design engineer in obtaining an optimum relation between the generator stator winding and the rotor damper circuit.

To attest the validity and, therefore, the usefulness of the curves of Figs. 1 and 3, a tabulation of test and curve values for various aircraft a-c generators is shown in Table I. The values for voltage unbalance from Figs. 1 and 3 are slightly higher than the test values. This is accountable for from the fact that the test value for Z_2 to Z_0 are obtained from tests conducted in accordance with procedures[6] which do not yield these impedances of the saturated and, consequently, lower values of the unbalanced load test conditions.

Conclusions

1. A simplified method of solution for the terminal voltages of a regulated, 3-phase synchronous generator carrying load on one phase only has been developed and verified. The solution is such that for a

defined load current the negative-sequence and zero-sequence impedance values are the only variables.

2. Curves of per-cent voltage unbalance for various values of negative-sequence and zero-sequence impedance are constructed for the 2/3-rated and 1/3-rated single-phase line-current conditions of the military specifications.[3,4] These curves emphasize the importance of the negative-sequence impedance in the reduction of voltage unbalance but also show that the zero-sequence impedance is a factor. The machine designer can use these results in the selection of generator characteristics for optimum design.

3. Curves of per-cent voltage unbalance for various values of negative-sequence reactance and resistance are constructed for the 2/3-rated single-phase line-current conditions of the military specifications.[3,4] These curves show that the negative-sequence resistance, as well as the negative-sequence reactance, plays an important roll in the reduction of voltage unbalance. The machine designer can apply these results to the evaluation of the rotor damper circuit for a particular design.

Appendix. Development of the Equations of the Terminal Voltages on a 3-Phase Synchronous Generator Carrying a Line-to-Neutral Load on One Phase Only

By writing Kirchhoff's equations in terms of the symmetrical component quantities, the following fundamental relations are obtained

$$I_a = I \tag{4}$$

$$I_b = 0 \tag{5}$$

$$I_c = 0 \tag{6}$$

$$I_{a0} = I_{b0} = I_{c0} = \frac{(I_a + I_b + I_c)}{3} = \frac{I_a}{3} = \frac{I}{3} \tag{7}$$

$$I_{a1} = \frac{(I_a + aI_b + a^2I_c)}{3} = \frac{I_a}{3} = \frac{I}{3} \tag{8}$$

$$I_{b1} = \frac{(I_b + aI_c + a^2I_a)}{3} = \frac{a^2I_a}{3} = \frac{a^2I}{3} \tag{9}$$

$$I_{c1} = \frac{(I_c + aI_a + a^2I_b)}{3} = \frac{aI_a}{3} = \frac{aI}{3} \tag{10}$$

$$I_{a2} = \frac{(I_a + a^2I_b + aI_c)}{3} = \frac{I_a}{3} = \frac{I}{3} \tag{11}$$

$$I_{b2} = \frac{(I_b + aI_a + a^2I_c)}{3} = \frac{aI_a}{3} = \frac{aI}{3} \tag{12}$$

$$I_{c2} = \frac{(I_c + a^2I_a + aI_b)}{3} = \frac{a^2I_a}{3} = \frac{a^2I}{3} \tag{13}$$

The generator internal drops are

$$V_{a0} = Z_{a0}I_{a0} = \frac{Z_{a0}I}{3} \tag{14}$$

$$V_{a1} = Z_{a1}I_{a1} = \frac{Z_{a1}I}{3} \tag{15}$$

$$V_{a2} = Z_{a2}I_{a2} = \frac{Z_{a2}I}{3} \tag{16}$$

$$V_{b0} = Z_{b0}I_{b0} = \frac{Z_{b0}I}{3} \tag{17}$$

$$V_{b1} = Z_{b1}I_{b1} = \frac{a^2Z_{b1}I}{3} \tag{18}$$

$$V_{b2} = Z_{b2}I_{b2} = \frac{aZ_{b2}I}{3} \tag{19}$$

$$V_{c0} = Z_{c0}I_{c0} = \frac{Z_{c0}I}{3} \tag{20}$$

$$V_{c1} = Z_{c1}I_{c1} = \frac{aZ_{c1}I}{3} \tag{21}$$

$$V_{c2} = Z_{c2}I_{c2} = \frac{a^2Z_{c2}I}{3} \tag{22}$$

Assuming that the generated voltages are equal, balanced, and 120 degrees apart

$$E_a = V_a + V_{a0} + V_{a1} + V_{a2} = E \tag{23}$$

$$E_b = V_b + V_{b0} + V_{b1} + V_{b2} = a^2E \tag{24}$$

$$E_c = V_c + V_{c0} + V_{c1} + V_{c2} = aE \tag{25}$$

Combining equations 14 through 25

$$E = V_a + \frac{I}{3}(Z_{a0} + Z_{a1} + Z_{a2}) \tag{26}$$

$$a^2E = V_b + \frac{I}{3}(Z_{b0} + a^2Z_{b1} + aZ_{b2}) \tag{27}$$

$$aE = V_c + \frac{I}{3}(Z_{c0} + aZ_{c1} + a^2Z_{c2}) \tag{28}$$

Assuming that the sequence impedances of each phase are equal, equations 26, 27, and 28 become

$$E = V_a + \frac{I}{3}(Z_0 + Z_1 + Z_2) \tag{29}$$

$$a^2E = V_b + \frac{I}{3}(Z_0 + a^2Z_1 + aZ_2) \tag{30}$$

$$aE = V_c + \frac{I}{3}(Z_0 + aZ_1 + a^2Z_2) \tag{31}$$

Multiplying equation 29 by a^2, equating equations 29 and 30, and solving for V_b gives

$$V_b = a^2V_a + \frac{I}{3}[(a^2 - 1)Z_0 + (a^2 - a)Z_2] \tag{32}$$

Multiplying equation 29 by a, equating equations 29 and 31, and solving for V_c gives

$$V_c = aV_a + \frac{I}{3}[(a - 1)Z_0 + (a - a^2)Z_2] \tag{33}$$

References

1. Methods for Prediction of Steady-State Performance for Unbalanced Regulated 3-Phase Generators, B. J. Wilson. *AIEE Transactions*, vol. 72, pt. II, 1953 (Jan. 1954 section), pp. 413–422.

2. Calculations on Voltage Unbalance for 3-Phase Synchronous Systems, B. J. Wilson, W. K. Gardner. *Ibid.*, vol. 73, pt. II, 1954 (Jan. 1955 section), pp. 426–37.

3. Electrical Power, Aircraft Characteristics of. *Military Specification MIL-E-7894*, U. S. Bureau of Aeronautics, Washington, D. C., April 25, 1952, sect. 4.1.1, and 4.1.3.

4. Generators and Regulators; Aircraft, Alternating Current, General Specification for. *Military Specification MIL-G-6099*, U. S. Bureau of Aeronautics, Washington, D. C., April 19, 1950, sect. 4.5.7, and 4.5.7.1.

5. Machine Characteristics, C. F. Wagner. *Electrical Transmission and Distribution Reference Book*, Westinghouse Electric Corporation, East Pittsburgh, Pa., 1950, chap. 6, pp. 145–94.

6. Test Code for Synchronous Machines. *AIEE Standard No. 503*, June 1945.

Discussion

B. J. Wilson and W. K. Gardner (Naval Research Laboratory, Washington, D. C.): We credit the author for his ingenuity in treating voltage unbalance within the limits that permit prediction of regulator action. The method appears useful for general unbalanced loading on wye-wye neutral connected, wye-wye, and wye-delta circuits.

The author has restricted his application to single-phase loading in wye-wye neutral connected circuits, but we would like to amplify its use. It works to advantage even for general 3-phase unbalanced loading.

When applying the method to 3-phase unbalanced loading of the wye-wye neutral connected circuit, equations 34, 35, and 36 form a convenient starting point. (These were derived from equations 7, 4, and 5 of reference 2 of the paper.)

$$0 = V_a\left[\frac{Z_0}{Z_a} + 1\right] + V_b\left[\frac{Z_0}{Z_b} + 1\right] + V_c\left[\frac{Z_0}{Z_c} + 1\right] \tag{34}$$

$$3E = V_a\left[\frac{Z_1}{Z_a} + 1\right] + V_b(a)\left[\frac{Z_1}{Z_b} + 1\right] + V_c(a^2)\left[\frac{Z_1}{Z_c} + 1\right] \tag{35}$$

$$0 = V_a\left[\frac{Z_2}{Z_a} + 1\right] + V_b(a^2)\left[\frac{Z_2}{Z_c} + 1\right] +$$
$$V_c(a)\left[\frac{Z_2}{Z_c} + 1\right] \tag{36}$$

Equations 37, 38, and 39 are in suitable form for application of the author's method, where V_a has been chosen for reference and equations 34 and 35 selected for use and rearranged.

$$V_a = \text{reference terminal voltage} \tag{37}$$

$$V_b\left[\frac{Z_0}{Z_b} + 1\right] + V_c\left[\frac{Z_0}{Z_c} + 1\right] = -V_a\left[\frac{Z_0}{Z_a} + 1\right] \tag{38}$$

$$V_b(a^2)\left[\frac{Z_2}{Z_b} + 1\right] + V_c(a)\left[\frac{Z_2}{Z_c} + 1\right]$$

$$= -V_a\left[\frac{Z_2}{Z_a}+1\right] \quad (39)$$

1. Assume a value for V_a.

2. Obtain corresponding values of V_b and V_c from simultaneous solution of equations 38 and 39 where knowledge is assumed of the values of machine sequence impedances Z_0, Z_2 and the phase load impedances Z_a, Z_b, Z_c.

3. Adjust the voltage level of the values of V_a, V_b, and V_c in a manner appropriate to the method of regulator sensing.

Similar equations, though more complicated, may be readily developed for the most general case of both self- and mutual impedance unbalanced three-phase loads.

When applying the method to the wye-wye circuit, equations 40–43 form the basic starting point.

$$3E=\left[\frac{1}{Z_a}\right][Z_1+Z_a-a^2(Z_1+Z_c)]\,V_a+$$
$$\left[\frac{1}{Z_b}\right][aZ_1-Z_b-a^2(Z_1+Z_b+Z_c)]\,V_b$$
$$\quad (40)$$

$$0=\left[\frac{1}{Z_a}\right][Z_2+Z_a-a(Z_2+Z_c)]\,V_a+$$
$$\left[\frac{1}{Z_b}\right][a^2Z_2-Z_b-a(Z_2+Z_b+Z_c)]\,V_b$$
$$\quad (41)$$

or

$$V_b=\left[-\frac{Z_b}{Z_a}\right]\left[\frac{Z_2+Z_a-a(Z_2+Z_c)}{a^2Z_2-Z_b-a(Z_2+Z_b+Z_c)}\right]$$
$$[V_a]$$

$$V_c=[-Z_c]\left[\frac{V_a}{Z_a}+\frac{V_b}{Z_b}\right] \quad (42)$$

$$V_{ab}=V_a-V_b;\quad V_{bc}=V_b-V_c;$$
$$V_{ca}=V_c-V_a \quad (43)$$

1. Assume a value of V_a, the reference terminal voltage $(L-N)$.

2. Calculate corresponding values of V_b and V_c from equations 41 and 42, respectively. Knowledge of the values of Z_2 and the phase load impedances Z_a, Z_b, Z_c is assumed.

3. Adjust the voltage level of V_a, V_b, V_c in a manner appropriate to the method of regulator sensing.

4. Calculate the final values of line-to-line voltages from equation 43.

When applying the method to the wye-delta connection, equations 44, 45, and 46 form the basic starting point.

$$(1-a^2)E=\left[\frac{Z_1}{Z_{ab}}-\frac{a^2Z_1}{Z_{ca}}+\frac{(1-a^2)}{3}\right]V_{ab}+$$
$$\left[\frac{aZ_1}{Z_{bc}}-\frac{a^2Z_1}{Z_{ca}}+\frac{(a-a^2)}{3}\right]V_{bc} \quad (44)$$

$$0=\left[\frac{Z_2}{Z_{ab}}-\frac{aZ_2}{Z_{ca}}+\frac{(1-a)}{3}\right]V_{ab}+$$
$$\left[\frac{a^2Z_2}{Z_{bc}}-\frac{aZ_2}{Z_{ca}}+\frac{(a^2-a)}{3}\right]V_{bc} \quad (45)$$

or

$$V_{bc}=\frac{\left[\dfrac{aZ_2}{Z_{ca}}-\dfrac{Z_2}{Z_{ab}}+\dfrac{(a-1)}{3}\right]}{\left[\dfrac{a^2Z_2}{Z_{bc}}-\dfrac{aZ_2}{Z_{ca}}+\dfrac{(a^2-a)}{3}\right]}V_{ab}$$

$$V_{ca}=-V_{ab}-V_{bc} \quad (46)$$

1. Assume a value of V_{ab}, the reference terminal voltage (line-to-line).

2. Calculate the corresponding values of V_{bc} and V_{ca} from equations 45 and 46, respectively. Knowledge of the values of Z_2 and the delta load impedances Z_{ab}, Z_{bc}, Z_{ca} is assumed.

3. Adjust the voltage level of the values of V_{ab}, V_{bc}, V_{ca} in a manner appropriate to the method of regulator sensing.

We wish to emphasize that the general synchronous system equations are not independent of Z_1. The author has derived equations that are independent of Z_1 for a special purpose. In the general problem of system voltage control the parameter Z_1 must remain explicit along with excitation voltage. Also, we caution in the use of the proposed technique that the regulator behaves within the limits of knowledgeable regulator action.

T. F. Hardman: The discussers have made an appreciable supplement to the paper by extending the calculation method to general 3-phase unbalanced loading. Their contribution is appreciated.

The paper limited the method of calculation and the constructed curves to single-phase loading in wye-wye neutral connected circuits to demonstrate the effect of the machine impedances on the voltage unbalance performance as imposed by the unbalanced load requirements of the military specifications.[1]

I am in accord with the discussers' comment that the general synchronous system equations are not independent of Z_1. Z_1 does not occur in equations 2 and 3 of the paper because the phase current I is a defined quantity. The phase current I is actually dependent upon Z_1, as well as the other system impedances.

Knowledgeable regulator action was included in the paper to illustrate that the regulator does not behave idealistically. The method of adjusting the voltage level in accordance with the method of regulator sensing, used by the author as well as the discussers, produces accurate results only when the regulator behaves exactly.

REFERENCE

1. See references 3 and 4 of the paper.

Sensing Methods Applicable to a 3-Phase Load Transfer Contactor

R. W. STINEMAN
ASSOCIATE MEMBER AIEE

P. L. EPSTEIN
NONMEMBER AIEE

A LOAD transfer contactor is employed in an electric circuit as illustrated in Fig. 1. As is evident from the figure, the purpose of the contactor is to connect a bank of loads to one of two possible power sources. While the contactor could be either manually or automatically controlled, only automatically controlled self-contained contactors are considered in this paper. Also, the contactor could be employed in either a single-phase or a 3-phase circuit, but this paper deals only with 3-phase circuits.

Load transfer contactors conventionally employed in aircraft 3-phase electric systems generally contain three sensing relays, each relay being connected from a phase of the normal power source to neutral. In the event that one of the sensing relays drops out as a result of a low line-to-neutral voltage, the load is transferred to the alternate power source. Power to operate the main contacts is drawn from the alternate power source. In a 200/115-volt system, the sensing relays are commonly adjusted to pick up under 100

volts and to drop out over 80 volts. Such a sensing circuit has a number of evident shortcomings:

1. If there is a line-to-line fault on the normal power feeder, the contactor may not respond properly.

2. The low ratio of pickup to dropout voltage means that each sensing relay operates with relatively little contact pressure and is therefore vulnerable to aircraft vibration.

3. The relatively close percentage tolerance on pickup and dropout voltage creates a difficult problem of compensating for environmental conditions such as temperature variations.

4. Loss of one phase of the normal power source may initiate a transfer even though the alternate power source has also lost a phase and is therefore not a better source of power. This factor is most significant in a power system which is arranged for either paralleled or isolated operation of its alternators. During paralleled operation in such a system, a fault may affect equally
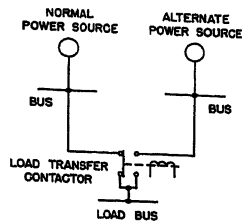
both the normal and alternate power sources and cause a momentary useless operation of the transfer contactor.

An obvious solution to the first of these items to to install three additional sensing relays connected line to line. However, the use of a total of six relays with their associated circuits makes the contactor quite bulky and heavy and, of course, the last three of the foregoing disadvantages still prevail. Furthermore, failure of any one of the six relays would cause a false transfer. It thus appears that a fresh approach to the problem of sensing circuitry is in order.

Fundamental Concepts

Substantial improvement over previous practice has been attained in contactors currently under development through the application of certain fundamental concepts new to the industry. The first of these concepts is that of comparative rather than absolute sensing. By this it is meant that the sensing circuit is designed to compare the voltage of the normal power source to the voltage of the alternate power source as a reference, rather than to refer the normal source voltage to an arbitrary fixed reference. In essence, comparative sensing is the true objective of a transfer contactor, in that it is desired to supply the loads with voltage as nearly normal as possible. One of the fruits of comparative sensing is the elimination of the second and third shortcomings listed, for the older sensing circuit. This is best explained by a numerical example. If the line-to-neutral voltages of two power sources were directly compared, it would be found that the sensing relay would be required to pick up under 35 volts and drop out over

15 volts in order to have performance comparable to the older system. Thus, the ratio of pickup to dropout voltage is considerably increased and a wider percentage tolerance on relay operation is available. Comparative sensing also eliminates the fourth objection to the older type of sensing circuit.

A second new concept is the employment of symmetrical components of normal and alternate power source voltage. It may be recalled that the positive-sequence component of three-phase voltages indicates the general level of voltage, while the negative- and zero-sequence components indicate the degree of unbalance of the voltages. However, no zero-sequence voltage is developed during a line-to-line fault and since an objective is to detect line-to-line faults, it is apparent that zero-sequence voltage is of little interest. On the other hand, negative-sequence voltage is developed sensing during any unsymmetrical fault; so by negative-sequence voltage, it is possible to detect both line-to-neutral and line-to-line faults. If such sensing is employed, it is necessary also to sense positive-sequence voltage in order to detect 3-phase faults or other failures which equally affect all phases of the normal power source.

Calculations were carried out to determine the magnitude of positive- and negative-sequence voltage during resistive line-to-neutral and line-to-line faults at the terminals of a Westinghouse type-8QL60A 60-kva alternator. The calculations were simplified to the extent that saturation, saliency, and limitations on maximum excitation were neglected. The results of the calculations are plotted in Fig. 2 as a function of fault voltage. These curves apply with little error to any load within the rating of the alternator. As a result of voltage regulator action, the magnitude of the positive-sequence voltage is relatively unaffected by a line-to-neutral or a line-to-line fault.

Combining the concepts of comparative and sequence voltage sensing, the following definitions are derived

$$E_{1d} = E_{1a} - E_{1n}$$

where

E_{1d} = differential positive-sequence voltage
E_{1a} = positive-sequence voltage of the alternate power source
E_{1n} = positive-sequence voltage of the normal power source

$$E_{2d} = E_{2n} - E_{2a}$$

where

E_{2d} = differential negative-sequence voltage
E_{2n} = negative-sequence voltage of the normal power source
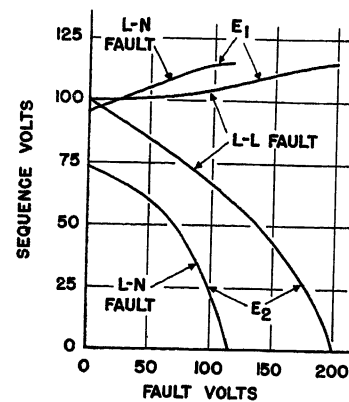
Fig. 2. Positive- and negative-sequence voltages during an unsymmetrical short circuit in an aircraft electrical system

E_{2a} = negative-sequence voltage of the alternate power source

Further

$$E_d = E_{1d} + E_{2d}$$

where E_d is called the differential sequence voltage.

There are three general conditions to which a load transfer contactor should respond:

1. Complete failure of the normal power source.

2. A partial but symmetrical failure of the normal power source as might be occasioned by a resistive 3-phase fault or insufficient alternator excitation.

3. A partial unsymmetrical failure of the normal power source such as an unsymmetrical short circuit or open circuit on the normal power feeder.

The first condition will be detected if the remaining conditions are detected, and need be given no special consideration.

Fig. 3. A load transfer contactor employing sequence-filter sensing circuits

Stineman, Epstein—Sensing Methods for a Load Transfer Contactor

Fig. 4. A simplified block diagram of the contactor illustrated in Fig. 3   Fig. 6. Outputs of a sensing circuit of the type illustrated in Fig. 7
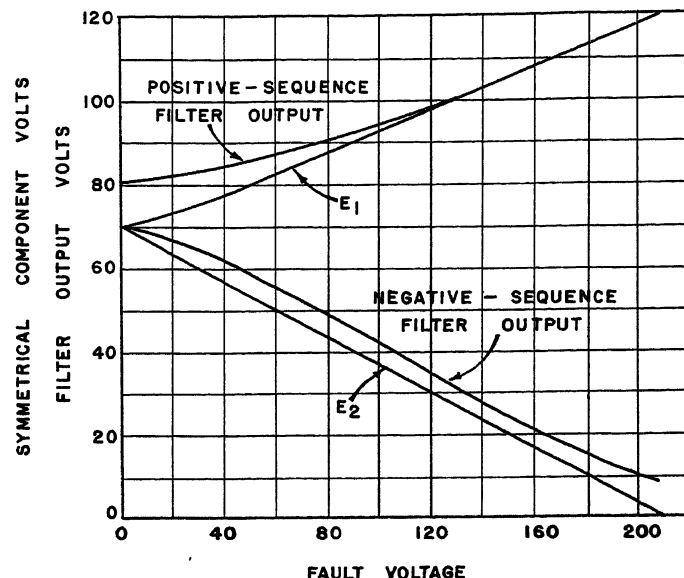
The second condition determines the required sensitivity to differential positive-sequence voltage; in line with past practice, a pickup-dropout range of 35 to 15 volts would appear appropriate. The third condition determines the required sensitivity to differential negative-sequence voltage. Referring to Fig. 2, it is evident that a sensitivity of 50 to 30 volts would result in a response comparable to the older type of sensing for a line-to-neutral fault. However, in the interest of standardization of sensing circuitry, a sensitivity of 40 to 20 volts was selected for both differential positive-sequence voltage and differential negative-sequence voltage or any equivalent combination. In other words, the contactor is made sensitive to simply a differential sequence voltage range of 40 to 20 volts, regardless of the combination of positive- and negative-sequence voltages.

Another factor which should be mentioned is the matter of time delay; i.e., the contactor should be able to override certain system transients. The older style sensing circuit was required to override the following conditions:

Fig. 5. Detailed circuit of the type of sequence filter used in Fig. 4. With normal phase sequence, abc, the filter operates as a negative-sequence filter

1. Application of load to the normal power source.

2. Faults on the load side of the contactor. This feature is necessary to allow such a fault to be cleared by circuit protectors in the normal power feeder before the fault is interrupted by the contactor itself, a phenomenon which leads to a cycling condition, which tends to prolong the duration of the fault, and which imposes upon the main contacts the severe burden of switching the fault current.

With comparative sensing, two additional conditions must be overriden by the contactor time delay:

3. Removal of load from the alternate power source.

4. Overvoltage of the alternate power source caused by faulty regulating equipment.

For conditions 1, 3, and 4, a delay of approximately 0.15 second appears to be adequate in most cases. The delay required for condition 2 depends on the type of fault protection employed. The contactors described in the succeeding sections of this paper were used in an application where a delay of 0.35 to 0.60 second was required.

Another characteristic which must be controlled in the design of a load transfer contactor is the difference between the transfer and retransfer voltages. It should be realized that when a transfer contactor operates, the system in which it is installed is altered to some extent. Consequently, it is not uncommon for the voltage at the normal power source terminals to increase slightly after the transfer operation. If the pickup and dropout voltages of the sensing elements are too close together, a cycling condition

may occur in which the contactor repeatedly transfers and retransfers. It is therefore necessary that the spread between transfer and retransfer voltage be at least 7 differential sequence volts, and it is desirable that the spread be at least 10 volts.

Contactor Employing Sequence-Filter Sensing Circuits

A straightforward application of the foregoing principles leads to the load transfer contactor illustrated in Fig. 3 and for which a simplified block diagram is shown in Fig. 4. Note that both the alternate and normal power source voltages are applied to the inputs of both positive- and negative-sequence filters. The output of each filter is then rectified and applied to one coil of a 4-coil polarized relay. The polarity of the connections to the relay is such that the relay, in ef-
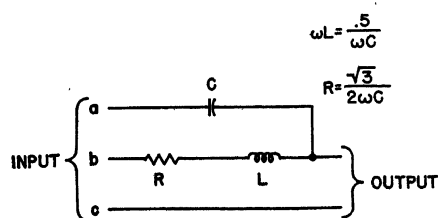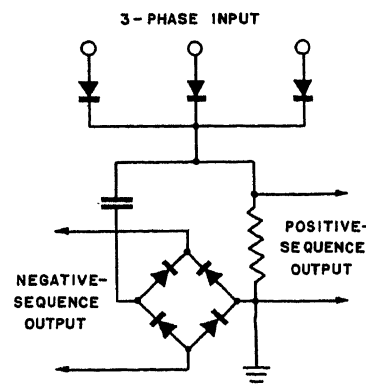
Fig. 7. A sensing circuit based on a 3-phase rectifier. This circuit is relatively insensitive to frequency and temperature changes
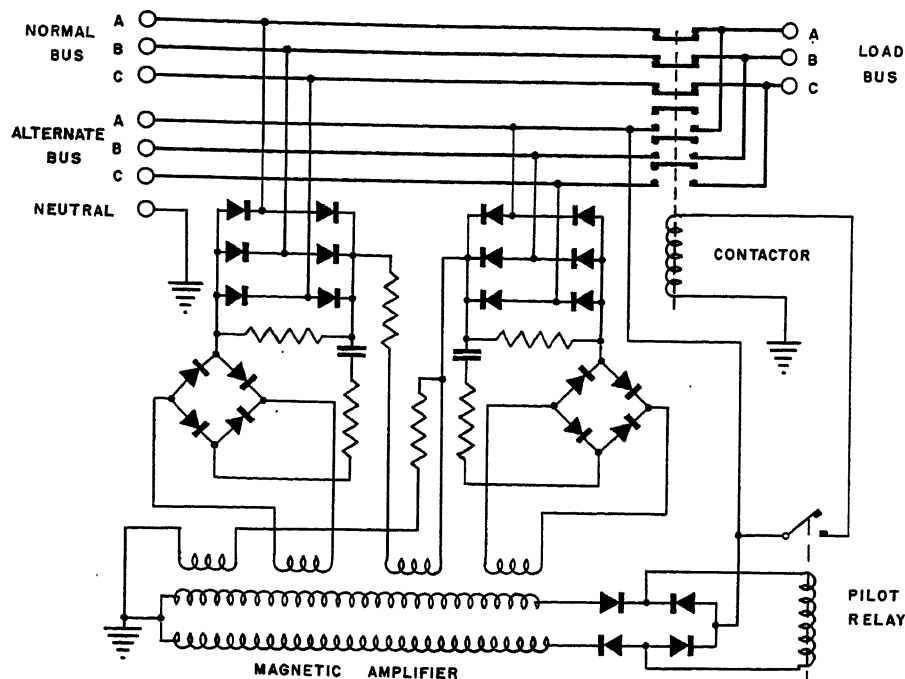
Fig. 8. Circuit diagram of a load transfer contactor employing sensing circuits of the type shown in Fig. 7. A magnetic amplifier compares the outputs of the sensing circuits and incorporates necessary time delay

fect, responds to differential sequence voltage as has been defined. When the differential sequence voltage reaches a value of 40 volts, the polarized relay operates to actuate the time-delay relay which in turn energizes the main contactor with power supplied by the alternate power source. Thus the loads are transferred to the alternate power source. If the differential sequence voltage subsequently drops to under 20 volts, the polarized relay drops out, thus causing the loads to be retransferred to the normal power source.

The design of the sequence filter circuits is shown in detail in Fig. 5. The filter is shown connected as a negative-sequence filter. However, exactly the same circuit could be used as a positive-sequence filter simply by interchanging

any two of the input leads, a feature which is characteristic of positive- and negative-sequence filters. The output of the filter is proportional to the sequence voltage being sensed. Other filter circuits would be possible but the one illustrated appears to be best suited to this application.

The contactor employing sequence filters was tested and its performance was found to be essentially as expected. A desirable characteristic of the circuit, that was discovered during the test program, was that the contactor would not connect the loads to a power source whose phase sequence had been reversed. The principal disadvantage of the contactor was that the sequence filters were found to be sensitive to frequency changes. As a result, the operating points did not remain within the desired limits at all combinations of operating temperature and frequency.

The weight of the contactor, including the sensing circuitry, is about 12 pounds. The main contacts are rated 150 amperes per phase at 200/115 volts, 400 cycles per second.

An Improved Transfer Contactor

The validity of a transfer contactor based on symmetrical component sensing having been established, the development of an improved unit was undertaken. The first difficulty to be attacked was the frequency and temperature sensitivity of the sequence filters, alluded to in connec-

tion with the tuned filter of Fig. 5. While other filters employing reactive elements might be designed, all filters of this type suffer from frequency sensitivity. Deviations of system frequency from the design center value of 400 cycles per second do not affect the positive-sequence filter very seriously, but the normal variation of aircraft system frequency from 380 to 420 cycles per second will result in a spurious output equivalent to 4 negative-sequence volts from the filter of Fig. 5 at the extreme limits of the frequency range. This is a substantial portion of the 20-to-40 volt figure which is to be monitored. In addition, networks having reactive elements are sensitive to temperature changes. While this obstacle is not inherent in the circuits, as frequency sensitivity is, it is practically very difficult to secure temperature-compensated reactive elements sufficiently good to avoid another 4 volts or so of spurious output at some point in the normal ambient temperature range of minus 60 to plus 85 degrees centigrade.

In view of the afore-mentioned difficulties, it was decided to make an attempt to eliminate the tuned filters. It was considered less important to synthesize the mathematically correct functions of the positive- and negative-sequence components than to have two functions which, in a general way, would manifest the average system voltage and the degree of unbalance between phases, provided these latter practical functions were reproducible and invariant with temperature and frequency. A possibility for eliminating the tuned positive-sequence filter lies in recognition of the fact that the output voltage of a 3-phase half-wave rectifier approximates closely the positive-sequence symmetrical component of the system. The approximation is very good unless
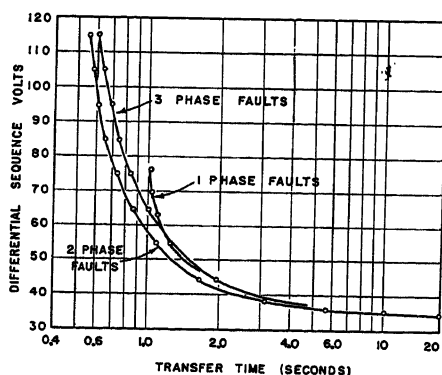
Fig. 9. Time-delay curves for the contactor of Fig. 8
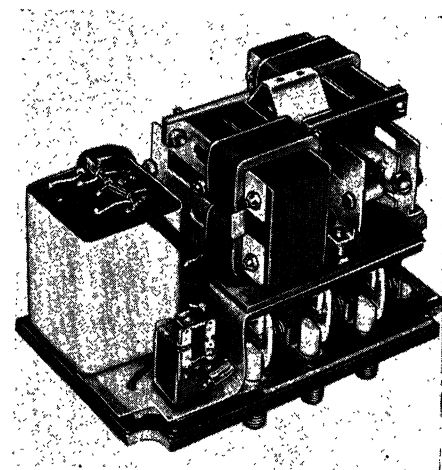
Fig. 10. Load transfer contactor in its finally developed configuration

there are large departures from the normal 120-degree phase relationship between the three line voltages. In the case of the negative-sequence component, it was considered that the ripple voltage produced by the 3-phase rectifier might be proportional to the desired function. Mathematical verification of this principle was carried out by means of a computer. Fig. 6 shows the agreement between the true positive- and negative-sequence voltages and the outputs of the rectifier unit, as calculated for a series of line-to-neutral faults on an actual a-c system. The agreement was considered sufficiently good for all practical purposes. By means of rectifier-type circuitry, frequency sensitivity has been eliminated from the filters, and the variation in output with temperature is substantially reduced. The basic circuit which emerged from this development is shown in Fig. 7. It should be noted that the circuit is no longer sensitive to the direction of phase rotation.

A practical circuit as used in the actual transfer contactor is shown in Fig. 8. Here the outputs of the positive- and negative-sequence filters of the normal and alternate power sources are fed into a magnetic amplifier which serves to combine differentially the positive- and negative-sequence signals of the normal and alternate power sources to provide an indication of which source is better. The amplifier output operates a pilot relay which transfers the contactor to the alternate source, should the latter become better than the normal source by more than 40 differential-sequence volts.

The power supply of the magnetic amplifier and that of the contactor are derived from the same phase of the alternate power source. This eliminates the possibility of the pilot relay demanding transfer when motor power might not be available to the contactor. The magnetic amplifier is operated underexcited and has high inductance signal windings so as to provide a time delay of up to 10 seconds to override the transient conditions discussed. The elimination of mechanical devices to produce this time delay has appreciably improved the reliability of the device.

The time delay provided by this arrangement is inversely proportional to the severity of the disturbance. Fig. 9 shows a plot of the transfer time versus the magnitude of the disturbance in terms of differential sequence volts. It will be noted that the results are slightly different for single-phase, 2-phase, and 3-phase-to-neutral faults, which would not be the case if the filter circuits developed voltages which were the exact mathematical analogues of the positive- and negative-sequence components. The chief source of this discrepancy is the output of the ripple filter which tends to be somewhat low at high values of negative-sequence voltages.

Fig. 10 shows the complete device. The contactor itself is a 3-pole double-throw unit, spring-loaded to the normal source. The unit shown is rated at 175 amperes per phase and weighs 7.5 pounds. Units having a 50-ampere rating have also been developed.

Conclusions

The application of new concepts in sensing circuit design has led to the following improved characteristics of load transfer contactors:

1. Reduction in weight and size.
2. Correct response to line-to-line as well as line-to-neutral faults.
3. Wider percentage tolerances in sensing elements with resulting increased ruggedness and improved performance.
4. Relative insensitivity to frequency and environmental changes.

Discussion

R. P. Sedgwick (Douglas Aircraft Company, Santa Monica, Calif.): The authors have made very clever use of a rectifier phenomenon which generally appears as a nuisance item and turned it into the solution of a frustrating problem. Their load transfer contactor evaluates between two power sources as to voltage magnitude and voltage balance and selects the better of the two for service to the loads. Overvoltage and frequency error are the only quantities not tested by the relay and these are generally controlled by other relays in a system.

I note that the transfer circuit may be more sensitive to short circuits on the load side of the contactor if the bus is supplied from a high-phase regulated generator, due to the reduction in positive phase sequence volts resulting from such regulation as compared with average sensing regulation.

Curiosity arises as to whether half-wave rectification for the negative sequence sensing circuit would give better sensitivity. It is noted that the circuit must be sensitive to the magnitude of the ripple voltage rather than its frequency, since the frequency (if the term is applicable) of the ripple voltage is "lower" when negative sequence voltage is present. Also I would like to know if the authors have investigated the behavior of the contactor for distorted bus voltage wave form as might be caused by certain types of loads.

I particularly like the statement concerning the unimportance of synthesizing the mathematically correct sequence functions.

R. W. Stineman and P. L. Epstein: Since our paper was prepared, the appearance of silicon rectifiers has made possible the design of a d-c solenoid to operate the main contacts. It is expected that this solenoid, together with the silicon rectifier, will occupy about the same space as the a-c solenoid shown in Fig. 10. The d-c design eliminates an objectionable hum which was produced by the a-c solenoid.

In reply to Mr. Sedgwick's comments, it is certainly true that the particular adjustment that was chosen for the contactor sensing elements may not be suitable for other applications. When high-phase sensing is employed in the system voltage regulators, a short circuit will produce less differential negative-sequence voltage and more differential positive-sequence voltage than is indicated by Fig. 2. However, the over-all differential sequence voltage should not be much different.

Full-wave rectification was employed in the sensing circuits (see Fig. 8) in order to eliminate, in so far as possible, the spurious negative-sequence signal resulting from the residual ripple in the rectifier output with balanced input. This is analogous, in principle, to striving for a high signal-to-noise ratio in communication circuits.

No particular attention was paid to wave form distortion, since the operating range of 20 to 40 differential sequence volts is much larger than any harmonic voltages present in the particular application discussed. It may be mentioned, however, that the "negative sequence" rectifier output will ignore harmonics which appear symmetrically on all three phases of the system.

ERRATA

In Table VI, page 330, the heading for column 6 now reads "NO TIF." This should be corrected to read "NONE," referring to no shunts in operation, while "TIF" indicates that the values given are for telephone influence factor. Likewise, the numerals in the headings for columns 3, 4, and 5 indicate the number of shunts in operation. A change also should be made in the RSS value in column 6 from 120, as it now appears, to 164.

Applications and Industry

Index for 1955

1. Technical Subject Index

478   *Technical Subjects*

2. Author Index